

(which can be regulable) of a gene in the cell. As used in general step (2), a biomolecule is a gene product (e.g., polypeptide, RNA, peptide or RNA oligonucleotide) of an exogenous gene -- a gene which has been introduced in the course of construction of the cell.

Biomolecules that bind to and alter the function of a candidate target are identified by various in vitro methods. Upon production of the biomolecule within a cell either in vitro or within an animal model system, the biomolecule binds to a specific site on the target, alters its intracellular function, and hence produces a phenotypic change (e.g. cessation of growth, cell death). When the biomolecule is produced in engineered pathogen cells in an animal model of infection, cessation of growth or death of the engineered pathogen cells leads to the clearing of infection and animal survival, demonstrating the importance of the target in infection and thereby validating the target.

#### **6.7.1.1. Comprising an exogenous regulable gene encoding the biomolecule**

A method for (1) identifying a biomolecule that produces a phenotypic effect on a cell (wherein the cell can be, for instance, a pathogen cell or a mammalian cell) and (2) simultaneous intracellular target validation, can comprise steps of introducing into an animal a cell comprising an exogenous regulable gene encoding the biomolecule, regulating expression of the gene to produce the biomolecule in the cell, and monitoring said cell in the animal for a phenotypic effect, compared to a suitable control cell. If the cell of this test manifests a phenotypic effect, this indicates that the biomolecule produced in the cell causes a phenotypic effect on the cell. If this phenotypic effect is the inhibition of growth of the cells, then the biomolecule can be termed a "biomolecular inhibitor" or a "biomolecular inhibitor of growth." It may be desirable to perform another test of intracellular function, using cell culturing techniques, wherein the cell comprising an exogenous regulable gene encoding the biomolecule of interest, and comprising the target cell component, is treated so as to turn on expression of the gene encoding the biomolecule, and one or more phenotypic characteristics of the cells in culture are monitored relative to suitable control cells, where the control cells do not produce the biomolecule. It may be preferable, where both "in culture" and "in animal" intracellular tests are performed, to do an "in culture" test first.

#### **6.7.1.2. Advantages of intracellular validation**

The purpose of intracellular validation for the combination of a potential target for drug action and molecule for drug development is two-fold. First, it demonstrates that the biomolecule under study produces a phenotypic effect on a living cell. In contrast with conditions in an in vitro binding test, the biomolecule in an intracellular test is exposed to a multitude of potential binding partners in the living cell, and interaction with one or more of these binding partners in the cell may be unproductive or result in undesirable effects. These effects are not detectable in an in vitro binding test. Second, where a biomolecule has been shown previously by in vitro tests to bind to a target cell component (that is, the biomolecule can be called a "biomolecular binder" of the target cell component, intracellular validation provides proof that the target cell component is essential to the maintenance of the original phenotype of the cell. Therefore, the target is validated for drug discovery and the biomolecule can then be utilized in a competitive binding assay to identify compounds that will have an effect on target molecule function.

Efficient binding between a biomolecule and a target cell component may be demonstrated in vitro; even binding, that inhibits activity of a target enzyme may be demonstrated in vitro. However, in the living cells, there could exist a redundant system that nullifies the effect of the biomolecule binding to the target cell component. For example, production of an enzyme having similar activity to that of the target cell component may be induced in the cells. By a mechanism such as this, the cell could escape any effect the biomolecule might otherwise cause by binding to the target cell component.

##### **6.7.1.2.1. Ensures that the target cell component is in its natural conformation as seen in the disease state**

Using an intracellular test to validate biomolecule/target cell component interaction is superior to using only an in vitro test using isolated molecules, because the intracellular test ensures that the target cell component is in its natural conformation and that the biomolecule "sees" the target cell component in that conformation, as that conformation occurs in a disease state. That in an intracellular test a biomolecule finds a site which ultimately causes a phenotypic effect on the cell indicates that the biomolecule is binding

to a functionally relevant site on the target cell component (e.g., an active site of an enzyme). Thus, molecules that are found to be structural analogs of the biomolecule and to compete with the biomolecule for a binding site on the target will also interact with the functionally relevant site of the target cell component, as functional analogs. A functional analog of the biomolecule can be found through competitive binding assays of the biomolecule against compounds (as in a library of compounds) that are potential binders of the target cell component. Structural analogs can also be found by rational drug design once a biomolecular binder is identified, by designing drugs that mimic the structure of the biomolecular binder. These structural analogs can be tested for their binding properties by techniques described herein.

**6.7.1.2.2. Biomolecule does not have to pass through a cell membrane or rely on inefficient uptake mechanisms of the cell**

A further advantage of the intracellular test in which the biomolecule is produced from one or more genes in the cell, is that the biomolecule does not have to pass through a cell membrane or rely on inefficient uptake mechanisms of the cell. Intracellular production of the biomolecule ensures that a biomolecule that interacts with a functional site on a target cell component to produce an effect will be detected, even if uptake of the biomolecule into the cell is limited. By the intracellular test, more biomolecules testing as being able to cause the desired phenotypic effect can be detected as candidates for further testing to find functional analogs for drug development. In a test employing extracellular addition of biomolecules. Biomolecules that bind the target cell component but are taken up by the cell only to a limited extent could be missed as candidates for further testing to find functional analogs for drug development. Limited uptake of a biomolecule which has been found to bind to a target in vitro is not necessarily a barrier to further steps towards drug development, as a structural portion of the ultimate compound to be administered as a drug can be selected for its stability, membrane solubility, efficient uptake, etc., and can be chemically combined with a compound whose structure mimics the active binding portion of the biomolecule. Intracellular production of the biomolecule, in an intracellular test of the effect of a biomolecule, as opposed to uptake from outside the cell, can also minimize degradation of the biomolecules from extracellular and intracellular degradative enzymes (e. g., proteases).

In further steps following one or more intracellular tests of the biomolecule/target cell component combination, one or more compounds that can also produce the phenotypic effect caused by the biomolecule can be identified in an in vitro competitive binding assay (which may be adapted for high-throughput screening) as compounds that compete with the biomolecule for a binding site on the target cell component. Target cell components can be isolated from the type of cell in which the phenotypic effect is desired (for instance, cells of pathogenic bacteria, yeast or fungi; mammalian cells, such as tumor cells), or from cells engineered to produce the cell component or a derivative of the cell component that would provide (at least some) structurally identical binding sites (e.g., a fusion protein). Compounds that produce the phenotypic effect observed with the biomolecule can be found in the competitive binding assay upon screening of libraries of compounds (for example, small molecule compounds or natural products or libraries that can be selected for having as their members compounds that have greater intracellular stability than biomolecules such as peptides or RNA oligonucleotides).

#### **6.7.1.3. Methods for identifying compounds that inhibit the growth of cells having a target cell component**

The invention includes methods for identifying compounds that inhibit the growth of cells having a target cell component. The target cell component can first be identified as essential to the growth of the cells in culture and/or under conditions in which it is desired that the growth of the cells be inhibited. These methods can be applied, for example, to various types of cells that undergo abnormal or undesirable proliferation, including cells of neoplasms (tumors or growths, either benign or malignant) which, as known in the art, can originate from a variety of different cell types. Such cells can be referred to, for example, as being from adenomas, carcinomas, lymphomas or leukemias. The method can also be applied to cells that proliferate abnormally in certain other diseases, such as arthritis, psoriasis or autoimmune diseases.

Described herein are similar methods for identifying inhibitors of target molecules or target cell components of pathogenic organisms. These methods can include a target validation procedure using an animal model for confirming that a cell component of a



pathogenic organism is essential, after infection with the organism has been established in a host, and that the inhibitor is effective against the organism after the organism has established the infection. A goal of the procedure is to identify compounds and/or gain the knowledge required to design compounds that can be used as antimicrobial agents to treat a human or other mammal having an infection of the organism.

#### **6.7.1.4. Methods for in vitro and in vivo validation of target and assay combinations**

The invention provides methods for in vitro and in vivo validation of target and assay combinations. Following selection of biomolecular binders to the isolated target cell component of interest, the invention can incorporate steps for (1) regulable (e.g., inducible) intracellular expression of a gene encoding the biomolecular binder and (2) monitoring cell viability in culture (e.g., cell growth in liquid media or agar plates) or in vivo (e.g., growth of introduced cells or pathogen virulence in an animal infection model) or both. If intracellular expression of the biomolecular binder inhibits the function of a target essential for growth (presumably by binding to the target at a biologically relevant site) cells monitored in step (2) will exhibit a slow growth or no growth phenotype. Targets found to be essential for growth by these methods are validated starting points for drug discovery, and can be incorporated into assays to identify more stable compounds that bind to the same site on the target as the biomolecule. Where the cells are pathogen cells and the desired phenotypic change to be monitored is inhibition of growth, the invention provides a procedure to examine the activity of target (pathogen) cell components in an animal infection model.

#### **6.7.1.5. Mimicking the environment for traditional antimicrobial therapy**

Controlled expression in cells of biomolecular binders to the target of interest mimics the environment for traditional antimicrobial therapy and validates targets as essential and appropriate for drug discovery. The technology facilitates choosing the best antimicrobial targets for drug discovery by facilitating, direct observation of the effect (phenotype) produced by target inhibition at a specific target subsite. The process is broadly applicable to a variety of targets. The process also validates target and biomolecular binder combinations as a direct path to high throughput screening for binding analogs of the

biomolecular binder, and is equally facile with targets that are gene products of genes of unknown function or genes of known function. Validated target and biomolecular binder assay combinations can be used directly in in vitro or in vivo competitive binding assays for screening chemical compound files. Compounds that compete with the biomolecular binders are identified as potential medicinal chemistry leads.

#### **6.7.1.6. Study as a target cell component a gene product of a particular cell type**

In the course of this method, it may be decided to study as a target cell component a gene product of a particular cell type (e.g., a type of pathogenic bacteria), wherein the target cell component is already known as being encoded by a characterized gene, as a potential target for a modulator to be identified. In this case, the target cell component can be isolated directly from the cell type of interest, assuming suitable culture methods are available to grow a sufficient number of cells, using methods appropriate to the type of cell component to be isolated (e.g., protein purification methods such as differential precipitation, ion exchange chromatography, gel chromatography, affinity chromatography, HPLC).

#### **6.7.1.7. Target cell component can be produced recombinantly**

Alternatively, the target cell component can be produced recombinantly, which requires that the gene encoding the target cell component be isolated from the cell type of interest. This can be done by any number of methods, for example known methods such as PCR, using template DNA isolated from the pathogen or a DNA library produced from the pathogen DNA, and using primers based on known sequences or combinations of known and unknown sequences within or external to the chosen gene. See, for example, methods described in "The Polymerase Chain Reaction," Chapter 15 of Current Protocols in Molecular Biology, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998. Other methods include cloning a gene from a DNA library (e.g., a cDNA library from a eucaryotic pathogen) into a vector (e.g., plasmid, phage, phagemid, virus, etc.) and applying a means of selection or screening, to clones resulting from a transformation of vectors (including a population of vectors now having inserted genes) into appropriate host cells. The screening method can take advantage of properties given to the host cells

by the expression of the inserted chosen gene (e.g., detection of the gene product by antibodies directed against it, detection of an enzymatic activity of the gene product), or can detect the presence of the gene itself (for instance, by methods employing nucleic acid hybridization). For methods of cloning genes in *E. coli*, which also may be applicable to cloning in other bacterial species, see, for example, "Escherichia coli, Plasmids and Bacteriophages," Chapter I of *Current Protocols in Molecular Biology*, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998. For methods applicable to cloning genes of eukaryotic origin, see Chapter 5 ("Construction of Recombinant DNA Libraries"), Chapter 9 ("Introduction of DNA Into Mammalian Cells") and Chapter 6 ("Screening of Recombinant DNA Libraries") of *Current Protocols in Molecular Biology*, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998.

Target proteins can be expressed with *E. coli* or other prokaryotic gene expression systems, or in eukaryotic gene expression systems. Since many eukaryotic proteins carry unique modifications that are required for their activities, e.g. glycosylation and methylation, protein expression can in some cases be better carried out in eukaryotic systems, such as yeast, insect, or mammalian cells that can perform these modifications. Examples of these expression systems have been reviewed in the following literature: *Methods in Enzymology*, Volume 185, eds D.V. Goeddel, Academic Press, San Diego, 1990; Geisse et al, *Protein Expression and Purification* 8:271-282, 1996; Simonsen and McGrogan, *Biologicals* 22: 85-94; Jones and Morikawa, *Current Opinions in Biotechnologies* 7: 512-516, 1996; Possee, *Current Opinions in Biotechnologies* 8:569-572.

Where a gene encoding a chosen target cell component has not been isolated previously, but is thought to exist because homologs of the gene product are known in other species, the gene can be identified and cloned by a method such as that used in Shiba et al., US 5,759,833, Shiba et al., US 5,629,188, Martinis et al., US 5,656,470 and Sassanfar et al., US 5,756,327. The teachings of these four patents are incorporated herein by reference in their entirety.

#### **6.7.1.8. Method should be used with target cell components which have not been previously isolated or characterized and whose functions are unknown**

It is an advantage of the target validation method that it can be used with target cell components which have not been previously isolated or characterized and whose functions are unknown. In this case, a segment of DNA containing an open reading frame (ORF; a cDNA can also be used, as appropriate to a eukaryotic cell) which has been isolated from a cell of a type that is to be an object of drug action (e.g., tumor cell, pathogen cell) can be cloned into a vector, and the target gene product of the ORF can be produced in host cells harboring the vector. The gene product can be purified and further studied in a manner similar to that of a gene product that has been previously isolated and characterized.

In some cases, the open reading frame (in some cases, cDNA) can be isolated from a source of DNA of the cells of interest (genomic DNA or a library, as appropriate), and inserted into a fusion protein or fusion polypeptide construct. This construct can be a vector comprising a nucleic acid sequence which provides a control region (e.g., promoter, ribosome binding site) and a region which encodes a peptide or polypeptide portion of the fusion polypeptide wherein the polypeptide encoded by the fusion vector endows the fusion polypeptide with one or more properties that allow for the purification of the fusion polypeptide. For example, the vector can be one from the pGEX series of plasmids (Pharmacia) designed to produce fusions with glutathione S-transferase.

#### **6.7.1.9. Host cells**

The isolated DNA having an open reading frame, whether encoding a known or an as yet unidentified gene product, when inserted into an expression construct, can be expressed to produce the target cell component in host cells. Host cells can be, for example, Gram-negative or Gram-positive bacterial cells such as *Escherichia coli* or *Bacillus subtilis*, respectively, or yeast cells such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* or *Pichia pastoris*. It is preferable that the target cell component to be used in target validation studies be produced in a host that is genetically related to the pathogen from which the gene encoding it was isolated. For example, for a Gram-negative bacterial pathogen, an *E. coli* host is preferred over a *Pichia pastoris* host. The target cell

component so produced can then be isolated from the host cells. Many protein purification methods are known that separate proteins on the basis of, for instance, size, charge, or affinity for a binding partner (e.g., for an enzyme, a binding partner can be a substrate or substrate analog), and these methods can be combined in a sequence of steps by persons of skill in the art to produce an effective purification scheme. For methods to manipulate RNA, see, for example, Chapter 4 in *Current Protocols in Molecular Biology* (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998.

An isolated cell component or a fusion protein comprising the cell component can be used in a test to identify one or more biomolecular binders of the isolated product (general step (1)). A biomolecular binder of a target cell component can be identified by in vitro assays that test for the formation of complexes of target and biomolecular binder noncovalently, bound to each other. For example, the isolated target can be contacted with one or more types of biomolecules under conditions conducive to binding, the unbound biomolecules can be removed from the targets, and a means of detecting bound complexes of biomolecules and targets can be applied. The detection of the bound complexes can be facilitated by having either the potential biomolecular binders or the target labeled or tagged with an adduct that allows detection or separation (e. g., radioactive isotope or fluorescent label; streptavidin, avidin or biotin affinity label).

Alternatively, both the potential biomolecular binders and the target can be differentially labeled. For examples of such methods see, e.g., WO 98/19162.

#### **6.7.1.10. Biomolecules to be tested and means for detection**

The biomolecules to be tested for binding to a target can be from a library of candidate biomolecular binders, (e.g., a peptide or oligonucleotide library). For example, a peptide library can be displayed on the coat protein of a phage (see, for examples of the use of genetic packages such as phage display libraries, Koivunen, E. et al., *J Biol. Chem.* 268:20205-20210 (1993)). The biomolecules can be detected by means of a chemical tag or label attached to or integrated into the biomolecules before they are screened for binding properties. For example, the label can be a radioisotope, a biotin tag, or a

fluorescent label. Those molecules that are found to bind to the target molecule can be called biomolecular binders.

#### **6.7.1.11. Fusion proteins**

An isolated target cell component, an antigenically similar portion thereof, or a suitable fusion protein comprising all of or a portion of or the entire target can be used in a method to select and identify biomolecules which bind specifically to the target. Where the target cell component comprises a protein, fusion proteins comprising all of, or a portion of, the target linked to a second moiety not occurring in the target as found in nature, can be prepared for use in another embodiment of the method. Suitable fusion proteins for this purpose include those in which the second moiety comprises an affinity ligand (e.g., an enzyme, antigen, epitope). The fusion proteins can be produced by the insertion of a gene encoding a target or a suitable portion of such gene into a suitable expression vector, which encodes an affinity ligand (e.g., pGEX-4T-2 and pET- 15b, encoding glutathione S-transferase and His-Tag affinity ligands, respectively). The expression vector can be introduced into a suitable host cell for expression. Host cells are lysed and the lysate, containing fusion protein, can be bound to a suitable affinity matrix by contacting the lysate with an affinity matrix under conditions sufficient for binding of the affinity ligand portion of the fusion protein to the affinity matrix.

##### **6.7.1.11.1. Fusion protein can be immobilized**

In one embodiment, the fusion protein can be immobilized on a suitable affinity matrix under conditions sufficient to bind the affinity ligand portion of the fusion protein to the matrix, and is contacted with one or more candidate biomolecules (e.g., a mixture of peptides) to be tested as biomolecular binders, under conditions suitable for binding of the biomolecules to the target portion of the bound fusion protein. Next, the affinity matrix with bound fusion protein can be washed with a suitable wash buffer to remove unbound biomolecules and non-specifically bound biomolecules. Biomolecules which remain bound can be released by contacting the affinity matrix with fusion protein bound thereto with a suitable elution buffer. Wash buffer can be formulated to permit binding of the fusion protein to the affinity matrix, without significantly disrupting binding of

specifically bound biomolecules. In this aspect, elution buffer can be formulated to permit retention of the fusion protein by the affinity matrix, but can be formulated to interfere with binding of the test biomolecule(s) to the target portion of the fusion protein. For example, a change in the ionic strength or pH of the elution buffer can lead to release of biomolecules, or the elution buffer can comprise a release component or components designed to disrupt binding of biomolecules to the target portion of the fusion protein.

Immobilization can be performed prior to, simultaneous with, or after contacting, the fusion protein with biomolecule, as appropriate. Various permutations of the method are possible, depending upon factors such as the biomolecules tested, the affinity matrix-ligand pair selected, and elution buffer formulation. For example, after the wash step, fusion protein with biomolecules bound thereto can be eluted from the affinity matrix with a suitable elution buffer (a matrix elution buffer, such as glutathione for a GST fusion). Where the fusion protein comprises a cleavable linker, such as a thrombin cleavage site, cleavage from the affinity ligand can release a portion of the fusion with the biomolecules bound thereto. Bound biomolecule can then be released from the fusion protein or its cleavage product by an appropriate method, such as extraction.

#### **6.7.1.12. Various methods to identify biomolecular binders**

One or more candidate biomolecular binders can be tested simultaneously. Where a mixture of biomolecules is tested, the biomolecules selected by the foregoing processes can be separated (as appropriate) and identified by suitable methods (e.g., PCR, sequencing, chromatography). Large libraries of biomolecules (e.g., peptides, RNA oligonucleotides) produced by combinatorial chemical synthesis or other methods can be tested (see e. a., Ohlmeyer, M.H.J. et al., *Proc. Natl. Acad. Sci. USA* 90:10922-10926 (1993) and DeWitt, S.H. et al., *Proc. Natl. Acad. Sci. USA* 90:6909-6913 (1993), relating to tagged compounds; see also Rutter, W.J. et al. U.S. Patent No. 5,010,175; Huebner, V.D. et al., U.S. Patent No. 5,182,366; and Geysen, H.M., U.S. Patent No. 4,833,092). Random sequence RNA libraries (see Ellington, A.D. et al., *Nature* 346:818-822 (1990); Bock, L.C. et al., *Nature* 355:584-566 (1992); and Szostak, J.W., *Trends in Biochem. Sci.* 17:89-93 (March, 1992)) can also be screened according to the present method to select

RNA molecules which bind to a target. Where biomolecules selected from a combinatorial library by the present method carry unique tags, identification of individual biomolecules by chromatographic methods is possible. Where biomolecules do not carry tags, chromatographic separation, followed by mass spectrometry to ascertain structure, can be used to identify individual biomolecules selected by the method, for example.

Other methods to identify biomolecular binders of a target cell component can be used. For example, the two-hybrid system or interaction trap is an *in vivo* system that can be used to identify polypeptides, peptides or proteins (candidate biomolecular binders) that bind to a target protein. In this system, both candidate biomolecular binders and target cell component proteins are produced as fusion proteins. The two-hybrid system and variations on it have been described (US 5,283,173 and US 5,468,614; Golemis, E.A. et al., pages 20.1.1-20.1.35 In *Current Protocols in Molecular Biology*, F.M. Ausubel et al., eds., John Wiley and Sons, containing supplements up through Supplement 40, 1997; two-hybrid systems available from Clontech, Palo Alto, CA).

Once one or more biomolecular binders of a cell component have been identified, further steps can be combined with those taken to identify the biomolecular binder, to identify those biomolecular binders that produce a phenotypic effect on a cell (where "a cell" can mean cells of a cell strain or cell line).

Thus, a method for identifying a biomolecule that produces a phenotypic effect on a first cell can comprise the steps of identifying a biomolecular binder of an isolated target cell component of the first cell, constructing a second cell comprising the target cell component and a regulable exogenous gene encoding the biomolecular binder, and testing the second cell for the phenotypic effect, upon production of the biomolecular binder in the second cell, where the second cell can be maintained in culture or introduced into an experimental animal. If the second cell shows the phenotypic effect upon intracellular production of the biomolecular binder, then a biomolecule that produces a phenotypic effect on the first cell has been identified. Testing the second cell is general step (2) of the invention, as the three general steps were outlined above.



#### **6.7.1.13. Host cells: Engineered to control expression**

Host cells (also, "second cells" in the terminology used above) of the cell type (e.g., species of pathogenic bacteria) the target was isolated from (or the gene encoding the target was originally isolated from, if the target is produced by recombinant methods), can be engineered to harbor a gene that can regulably express the biomolecular binder (e.g., under an inducible or repressible promoter). The ability to regulate the expression of the biomolecular binder is desirable because constitutive expression of the biomolecular binder could be lethal to the cell.

Therefore, inducible or regulated expression gives the researcher the ability to control if and when the biomolecular binder is expressed. The gene expressing the biomolecular binder can be present in one or more copies, either on an extra chromosomal structure, such as on a single or multicopy plasmid, or integrated into the host cell genome. Plasmids that provide an inducible gene expression system in pathogenic organisms can be used. For example, plasmids allowing tetracycline-inducible expression of a gene in *Staphylococcus aureus* have been developed.

#### **6.7.1.14. Genes for expression**

For intracellular expression of a biomolecule to be tested for its phenotypic effect in a eukaryotic cell (e.g., mammalian cell), the genes for expression can be carried on plasmid-based or virus-based vectors, or on a linear piece of DNA or RNA. For examples of expression vectors, see Hosfield and Lu, *Biotechniques*: 306-309, 1998; Stephens and Cockett, *Nucleic Acid Research* 17:7110, 1989; Wohlgemuth et al, *Gene Therapy*, 3:503-512, 1996; Ramirez-Solis et al, *Gene* 87:291-294, 1990, Dirks et al, *Gene* 149:387-388, 1994; Chenaalvala et al. *Current Opinion in Biotechnologies* 2:718-722, 1991; *Methods in Enzymology*, Volume 185, (D.V. Goeddel, ed.) Academic Press, San Diego, 1990. The genetic material can be introduced into cells using a variety of techniques, including whole cell or protoplast transformation, electroporation, calcium phosphate-DNA precipitation or DEAE- Dextran transfection, liposome mediated DNA or RNA transfer, or transduction with recombinant viral or retroviral vectors. Expression of the gene can be constitutive (e.g., ADHI promoter for expression in *S. cerevisiae* (Bennetzen, J.L. and Hall, B.D., J

Biol. Chem 257:3026-3031 (1982)), or CMV immediate early promoter and RSV LTR for mammalian expression) or inducible, as the inducible GAL I promoter in yeast (Davis, L.I. and Fink, G.R., Cell 61:965-978 (1990)). A variety of inducible systems can be utilized, for example, E. coli Lac repressor/operator system and Tn10 Tet repressor/operator systems have been engineered to govern regulated expression in organisms from bacterial to mammalian cells. Regulated gene expression can also be achieved by activation. For example, gene expression governed by HIV LTR can be activated by HIV or SIV Tat proteins in human cells; GAL4 promoter can be activated by galactose in a nonglucose-containing medium. The location of the biomolecule binder genes can be extra chromosomal or chromosomally integrated. The chromosome integration can be mediated through homologous or nonhomologous recombinations.

For proper localization in the cells, it may be desirable to tag the biomolecule binders with certain peptide signal sequences (for example, nuclear localization signal (NLS) sequences, mitochondria localization sequences). Secretion sequences have been well documented in the art.

#### **6.7.1.15. Fused biomolecular binders**

For presentation of the biomolecular binders in the intracellular system, they can be fused N-terminally, C-terminally, or internally in a carrier protein (if the biomolecular binder is a peptide), and can be fused (5', 3' or internally) in a carrier RNA or DNA molecule (if the biomolecular binder is a nucleic acid). The biomolecular binder can be presented with a protein or nucleic acid structural scaffold. Certain linkages (e.g., a 4-glycine linker for a peptide or a stretch of A's for an RNA) can be inserted between the biomolecular binder and the carrier proteins or nucleic acids.

In such engineered cells, the effect of this biomolecular binder on the phenotype of the cells can be tested, as a manifestation of the binding (implying binding to a functionally relevant site, thus, an activator, or more likely, an inhibitory) effect of the biomolecular binder on the target used in an in vitro binding assay as described above. An intracellular test can not only determine which biomolecular binders have a phenotypic effect on the cells, but at the same time can assess whether the target in the cells is essential for

maintaining the normal phenotype of the cells. For example, a culture of the engineered cells expressing a biomolecular binder can be divided into two aliquots. The first aliquot ("test" cells) can be treated in a suitable manner to regulate (e.g., induce or release repression of, as appropriate) the gene encoding the biomolecular binder, such that the biomolecular binder is produced in the cells. The second aliquot ("control" cells) can be left untreated so that the biomolecular binder is not produced in the cells. In a variation of this method of testing the effect of a biomolecular binder on the phenotype of the cells, a different strain of cells, not having a gene that can express the biomolecular binder, can be used as control cells. The phenotype of the cells in each culture ("test" and "control" cells grown under the same conditions, other than the expression of the biomolecular binder), can then be monitored by a suitable means (e.g., enzymatic activity, monitoring, a product of a biosynthetic pathway, antibody to test for presence of cell surface antigen, etc.). Where the change in phenotype is a change in growth rate, the growth of the cells in each culture ("test" and "control" cells grown under the same conditions, other than the expression of the biomolecular binder), can be monitored by a suitable means (e.g., turbidity of liquid cultures, cell count, etc.). If the extent of growth, or rate of growth of the test cells is less than the extent of growth or rate of growth of the control cells, then the biomolecular binder can be concluded to be an inhibitor of the growth of the cells, or a biomolecular inhibitor.

If the phenotype of the test cells is altered relative to that of the control cells, then the biomolecular binder can be concluded to be one that causes a phenotypic effect. In an optional additional test, isolated target cell component having a known function (e.g., an enzyme activity) can be tested for modulation of this known function in the presence of biomolecular binder under conditions conducive to binding of the biomolecular binder to the target cell component. Positive results in these tests should encourage the investigator to continue in the drug discovery process with efforts to find a more stable compound (than a peptide, polypeptide or RNA biomolecule) that mimics the binding properties of the biomolecular binder on the tested target cell component.

#### 6.7.1.16. Engineering strain of cells

A further test can, again, employ an engineered strain of cells that comprise both the target cell component and one or more genes encoding a biomolecule tested to be a biomolecular binder of the target cell component. The cells of the cell strain can be tested in animals to see if regulable expression of the biomolecular binder in the engineered cells produces an observable or testable change in phenotype of the cells. Both the "in culture" test for the effect of intracellular expression of the biomolecular binder and the "in animal" test (described below) for the effect of intracellular expression of the biomolecular binder can be applied not only towards drug discovery in the categories of antimicrobials and anticancer agents, but also towards the discovery of therapeutic agents to treat inflammatory diseases, cardiovascular diseases, diseases associated with metabolic pathways, and diseases associated with the central nervous system, for example.

Where the engineered strain of cells is a strain of pathogen cells or tumor cells, the object of the test is to see whether production of the biomolecular binder in the engineered strain inhibits growth of these cells after their introduction into an animal by the engineered pathogen. Such a test can not only determine which biomolecular binders are inhibitors of growth of the cells, but at the same time can assess whether the target in the cells is essential for maintaining growth of the cells (infection, for a pathogenic organism) in a host mammal. Suitable animals for such an experiment are, for example, mammals such as mice, rats, rabbits, guinea pigs, dogs, pigs, and the like. Small mammals are preferred for reasons of convenience.

The engineered cells are introduced into one or more animals ("test" animals) and into one or more animals in a separate group ("control" animals) by a route appropriate to cause symptoms of systemic or local growth of the engineered cells.

The route of introduction may be, for example, by oral feeding, by inhalation, by subdermal, intramuscular, intravenous, or intraperitoneal injection as appropriate to the desired result.

After the cell strain has been introduced into the test and control animals, expression of the gene encoding the biomolecular binder is regulated to allow production of the biomolecular binder in the engineered pathogen cells. This can be achieved, for instance, by administering to the test animals a treatment appropriate to the regulation system built into the cells, to cause the gene encoding the biomolecular binder to be expressed. The same treatment is not administered to the control animals, but the conditions under which they are maintained are otherwise identical to those of the test animals. The treatment to express the gene encoding the biomolecular binder can be the administration of an inducer substance (where expression of the biomolecular binder or gene is under the control of an inducible promoter) or the functional removal of a repressor substance (where expression of the biomolecular binder gene is under the control of a repressible promoter).

After such treatment, the test and control animals can be monitored for a phenotypic effect in the introduced cells. Where the introduced cells are constructed pathogen cells, the animals can be monitored for signs of infection (as the simplest endpoint, death of the animal, but also e.g., lethargy, lack of grooming behavior, hunched posture, not eating, diarrhea or other discharges; bacterial titer in samples of blood or other cultured fluids or tissues). In the case of testing engineered tumor cells, the test and control animals can be monitored for the development of tumors or for other indicators of the proliferation of the introduced engineered cells. If the test animals are observed to exhibit less growth of the introduced cells than the control animals, then the biomolecule can be also called a biomolecular inhibitor of growth, or biomolecular inhibitor of infection, as appropriate, as it can be concluded that the expression in vivo of the biomolecular inhibitor is the cause of the relative reduction in growth of the introduced cells in the test animals.

#### **6.7.2. In vitro assays**

Further steps of the procedure involve in vitro assays to identify one or more compounds that have binding and activating or inhibitory properties that are similar to those of the biomolecules which have been found to have a phenotypic effect, such as inhibition of growth. That is, compounds that compete for binding to a target cell component with the biomolecule would then be structural analogs of the biomolecules. Assays to identify such

compounds can take advantage of known methods to identify competing molecules in a binding assay. These steps comprise general step (3) of the method.

In one method to identify such compounds, a biomolecular inhibitor (or activator) can be contacted with the isolated target-cell component to allow binding, one or more compounds can be added to the milieu comprising the biomolecular inhibitor and the cell component under conditions that allow interaction and binding between the cell component and the biomolecular inhibitor, and any biomolecular inhibitor that is released from the cell component can be detected.

#### **6.7.2.1. Fluorescence**

One suitable system that allows the detection of released biomolecular inhibitor (or activator) is one in which fluorescence polarization of molecules in the milieu can be measured. The biomolecular inhibitor can have bound to it a fluorescent tag or label such as fluorescein or fluorescein attached to a linker.

Assays for inhibition of the binding of the biomolecular inhibitor to the cell component can be done in microtiter plates to conveniently test a set of compounds at the same time. In such assays, a majority of the fluorescently labeled biomolecular inhibitor must bind to the protein in the absence of competitor compound to allow for the detection of small changes in the bound versus free probe population when a compound which is a competitor with a biomolecular inhibitor is added (B.A. Lynch, et al., *Analytical Biochemistry* 247:77-82 (1997)). If a compound competes with the biomolecular inhibitor for a binding site on the target cell component, then fluorescently labeled biomolecular inhibitor is released from the target cell component, lowering the polarization measured in the milieu.

#### **6.7.2.2. Radioactive isotope**

In a further method for identifying one or more compounds that compete with a biomolecular inhibitor (or activator) for a binding site on a target cell component, the target cell component can be attached to a solid support, contacted with one or more

compounds, and contacted with the biomolecular inhibitor. One or more washing steps can be employed to remove biomolecular inhibitor and compound not bound to the cell component. Either the biomolecular inhibitor bound to the target cell component or the compound bound to the target cell component can be measured. Detection of biomolecular inhibitor or compound bound to the cell component can be facilitated by the use of a label on either molecule type, wherein the label can be, for instance, a radioactive isotope either incorporated into the molecule itself or attached as an adduct, streptavidin or biotin, a fluorescent label or a substrate for an enzyme that can produce from the substrate a colored or fluorescent product. An appropriate means of detection of the labeled biomolecular inhibitor or compound moiety of the biomolecular inhibitor- cell component complex or the compound-cell component complex can be applied. For example, a scintillation counter can be used to measure radioactivity. Radio labeled streptavidin or biotin can be allowed to bind to biotin or streptavidin, respectively, and the resulting complexes detected in a scintillation counter. Alkaline phosphatase conjugated to streptavidin can be added to a biotin-labeled biomolecular inhibitor or compound. Detection and quantitation of a biotin-labeled complex can then be by addition of pNPP substrate of alkaline phosphatase and detection by spectrophotometry, of a product which absorbs UV light at a wavelength of 405 nm. A fluorescent label can also be used, in which case detection of fluorescent complexes can be by a fluorometer. Models are available that can read multiple samples, as in a microtiter plate.

For example, in one type of assay, the method for identifying compounds comprises attaching the target cell component to a solid support, contacting the biomolecular inhibitor with the target cell component under conditions suitable for binding of the biomolecular inhibitor to the cell component, removing unbound biomolecular inhibitor from the solid support, contacting one or more compounds (e.g., a mixture of compounds) with the cell component under conditions suitable for binding of the biomolecular inhibitor to the cell component, and testing for unbound biomolecular inhibitor released from the cell component, whereby if unbound biomolecular inhibitor is detected, one or more compounds that displace or compete with the biomolecular inhibitor for a particular site on the target cell component have been identified.

Other methods for identifying compounds that are competitive binders with the biomolecule for a target can employ adaptations of fluorescence polarization methods. See, for instance, *Anal. Biochem.* 253(2):210-218 (1997), *Anal. Biochem.* 249(1):29-36 (1997), *BioTechniques* 17(3):585-589 (1994) and *Nature* 373:254-256 (1995).

Those compounds that bind competitively to the target cell component can be considered to be drug candidates. Further appropriate testing can confirm that those compounds which bind competitively with biomolecular inhibitors (or activators) possess the same activity as seen in an intracellular test of the effect of the biomolecular inhibitor or activator upon the phenotype of cells. Derivatives of these compounds having modifications to confer improved solubility, stability, etc., can also be tested for a desired phenotypic effect.

#### **6.7.3. Combining steps**

Combining steps for testing the phenotypic effects of a biomolecule, as can be produced in an intracellular test, with steps for identifying compounds that compete with the biomolecule for sites on a target cell component, yields a method for identifying a compound which is a functional analog of a biomolecule which produces a phenotypic effect on a cell. These steps can be to test, for the phenotypic effect, either in culture or in an animal model, or in both, a cell which produces a biomolecule by regulable expression of an exogenous gene in the cell, and to identify, if the biomolecule caused the phenotypic effect, one or more compounds that compete with the biomolecule for binding to a target cell component. If a compound is found to compete with the biomolecule for binding to the target cell component, then the compound is a functional analog of a biomolecule which produces a phenotypic effect on the cell. Such a functional analog can cause qualitatively a similar effect on the cell, but to a similar degree, lesser degree or greater degree than the biomolecule.

##### **6.7.3.1. Method for determining whether a target component of a cell is essential to producing a phenotypic effect on the cell**

A further embodiment of the invention combining general steps (1) and (2) is a method for determining whether a target component of a cell is essential to producing a phenotypic



effect on the cell, comprising isolating the target component from the cell, identifying a biomolecular binder of the isolated target component of the cell, constructing a second cell comprising the target component and a regulable, exogenous gene encoding the biomolecular binder, and testing the second cell in culture for an altered phenotypic effect, upon production of the biomolecular binder in the second cell, whereby, if the second cell shows the altered phenotypic effect upon production of the biomolecular binder, then the target component of the first cell is essential to producing the phenotypic effect on the first cell.

#### 6.7.3.1.1. Inhibit the proliferation of the cells

The methods described herein are well suited to the identification of compounds that can inhibit the proliferation of the cells of infectious agents such as bacteria, fungi and the like. In addition, a procedure such as the one outlined below can be used in the identification of compounds to inhibit the proliferation of cancer cells. The two procedures described below further illustrate the use of the methods described herein and would provide proof of principle of these methods with a known target for anticancer therapy.

Mammalian dihydrofolate reductase (DHFR) is a proven target for anticancer therapy. Methotrexate (MTX) is one of many existing drugs that inhibit DHFR. It is widely used for anticancer chemotherapy.

NIH 3T3 is a mouse fibroblast cell line that is able to develop spontaneous transformed cells when cultured in low concentration (2%) of calf serum in molecular, cellular and developmental biology medium 402 (MCDB) (M. Chow and H. Rubin, Proc. Natl. Acad. Sci. USA 95(8):4550-4555 (1998)). The transformed cells, which can be selectively inhibited by MTX (Chow and Rubin), are isolated.

Both the normal and transformed NIH3T3 cells are transfected with pTet- On plasmid (Clontech; Palo Alto, CA). Stable cell lines that express high levels of reverse tetracycline-control led activator (rtTA) are isolated and characterized for their normal or transformed phenotype (Chow and Rubin).

The DHFR gene (Genbank Accession # L26316) from the NIH 3T3 cell line is amplified by reverse transcription-PCR (RT-PCR) using poly A' RNA isolated from NIH 3T3 cells (Sambrook, J. et al., Molecular Cloning: A Laboratory Manual, 2nd edition, Cold Spring Harbor Laboratory Press, 1989). Active DHFR is expressed using the BacPAK Baculovirus Expression System (Clontech) or other appropriate systems. The expressed DHFR is purified and biotinylated and subjected to peptide binder identification as exemplified for bacterial proteins. The identified peptides are biochemically characterized for in vitro inhibition of DHFR activity. Peptides that inhibit DHFR are identified. A nucleic acid encoding each peptide can be cloned into a vector such as pGEX-4T2 (Pharmacia) to yield a vector which encodes a fusion polypeptide having the peptide fused to the N- terminus of GST. This can also be done by PCR amplification as exemplified herein for the peptide Pro-3. The fusion genes are cloned into plasmid pTRE (Clontech) for regulated expression. The constructed plasmid or the vector is cotransfected with pTK-Hyg into the stable NIH 3T3 cell line that expresses rtTA. The resulting cell lines, termed 3T3N-VITA (normal 3T3 cells that express rtTA and the DHFR inhibitory peptides), 3T3T-VITA (transformed 3T3 cells that express rtTA and the DHFR inhibitory peptides), or 3T3T-VITA control (transformed 3T3 cells that express rtTA and GST), are characterized for their normal or transformed phenotype (loss of contact inhibition, change in morphology, immortalization, etc. ).  $10^2$ - $10^1$  of 3T3T-VITA or 3T3T-VITA control cells are mixed with  $10^5$  3T3N-VITA and are grown in MCD 402 medium with 10% calf serum at 37°C for three days. Tetracycline is added to the medium to a final concentration of 0 to 1 ug/ml. In a control, 200 nM of MTX is added. The cultures are incubated for an additional eight days, and the number of foci formed are counted as described by M. Chow and H. Rubin, Proc. Natl. Acad. Sci. USA 95(8):4550-4555 (1998). Peptides that specifically inhibit foci formation of 3T3 transformed cells are identified.

A murine model of fibroblastoma (Kogerman, P. et al., Oncogene (12):1407-1416 (1997)) is used for evaluating the DHFR/peptide combination for identification of compounds for cancer therapy. Various amounts of 3T3T- VITA or 3T3T-VITA control cells ( $10^3$ ,  $10^4$ ,  $10^5$ ,  $10^6$  cells) are injected subcutaneously into 5 groups (10 in each group) of athymic nude mice (4-6 weeks old, 18-22 g) to determine the minimal dose needed for development of fibroblastomas in all of the tested animals. Upon determination of the minimal tumorigenic dose, 6 groups of athymic nude mice (10 each) are injected

subcutaneously (s.c.) with the minimal tumorigenic dose for 3T3T-VITA or 3T3T-VITA control cells to develop fibroblastoma. One week after injection, group I mice start receiving MTX s.c. at 2 mg/kg/day as positive control, group 2 to 5 start receiving 1, 2, 5, or 10 mg/kg/day of tetracycline, group 6 start receiving saline (vehicle) as control. Five weeks after the introduction of cells, all of the mice are sacrificed and tumors are removed from them. Tumor mass is measured and compared among the groups.

An effective peptide identified by these in vivo experiments can be used for screening libraries of compounds to identify those compounds that competitively bind to DHFR. One mechanism of tumorigenesis is overexpression of proto-oncogenes such as Ha-ras (Reviewed by Suarez, H.G., *Anticancer Research* 9(5):1331-1343 (1989)).

Compounds that inhibit the activities of the products of such proto- oncogenes can be used for cancer chemotherapy. What follows is a further illustration of the methods described herein, as applied to mammalian cells.

Transgenic mice that overexpress human Ha-ras have been produced. Such transgenic mice develop salivary and/or mammary adenocarcinomas (Nielsen, L.L. et al, *In Vivo* 8(5):1331-1343 (1994)). Secondary transgenic mice that express rtTA can be generated using the pTet-On plasmid from Clontech.

Human Ha-ras open reading frame cDNA (Genbank Accession #GO0277) is amplified by RT-PCR using polyA- RNA isolated from human mammary gland or other tissues. Active Ha-ras is expressed using the BacPAK Baculovirus Expression System (Clontech) or other appropriate systems. The expressed Ha-ras is purified and biotinylated and subjected to peptide binder identification as exemplified herein for bacterial proteins as target cell components. The identified peptides are biochemically characterized for in vitro inhibition of Ha-ras GTPase activity.

Peptides that inhibit Ha-ras are cloned into plasmid pTPE (Clontech) for regulated expression as an N-terminal fusion of GST. Such constructs are used to generate tertiary transgenic mice using the secondary transgenic mice. Transgenic mice that are able to

overexpress peptide genes are identified by Northern and Western analysis. Control mice that express GST are also identified.

Various doses of tetracycline are administered to the tertiary transgenic mice by s.c. or i.p. injection before or after tumor onset. Prevention or regression of tumors resulting from expression of the peptide genes are analyzed as described above for murine fibroblastoma. Peptides found to be effective in in vivo experiments will be used to screen compounds that inhibit human Ha-ras activity for cancer therapy.

#### **6.7.4. Disease targets**

The method of the invention can be applied more generally to mammalian diseases caused by: (1) loss or gain of protein function, (2) over- expression or loss of regulation of protein activity. In each case the starting point is the identification of a putative protein target or metabolic pathway involved in the disease. The protocol can sometimes vary with the disease indication, depending on the availability of cell culture and animal model systems to study the disease. In all cases the process can deliver a validated target and assay combination to support the initiation of drug discovery.

Appropriate disease indications include, but are not limited to, Alzheimer's, arthritis, cancer, cardiovascular diseases, central nervous system disorders, diabetes, depression, hypertension, inflammation, obesity and pain.

Appropriate protein targets putatively linked to disease indications include, but are not limited to (1) the leptin protein, putatively linked to obesity and diabetes; (2) a mitogen-activated protein kinase putatively linked to arthritis, osteoporosis and atherosclerosis; (3) the interleukin-1 beta converting protein putatively linked to arthritis, asthma and inflammation; (4) the caspase proteins putatively linked to neurodegenerative diseases such as Alzheimer's, Parkinson's and stroke, and (5) the tumor necrosis factor protein putatively linked to obesity and diabetes. Appropriate protein targets include also, but are not limited to, enzymes catalyzing the following types of reactions: (1) oxido-reductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases.

The arachidonic acid pathway constitutes one of the main mechanisms for the production of pain and inflammation. The pathway produces different classes of end products, including the prostaglandins, thromboxane and leukotrienes.

Prostaglandins, an end product of cyclooxygenase metabolism, modulate immune function, mediate vascular phases of inflammation and are potent vasodilators. The major therapeutic action of aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs) is proposed to be inhibition of the enzyme cyclooxygenase (COX). Anti-inflammatory potencies of different NSAIDs have been shown to be proportional to their action as COX inhibitors. It has also been shown that COX inhibition produces toxic side effects such as erosive gastritis and renal toxicity. The knowledge base regarding the toxic side effects of COX inhibitors has been gained through years of monitoring human therapies and human suffering. Two kinds of COX enzymes are now known to exist, with inhibition of COX 1 related to toxicity, and inhibition of COX2 related to reduction of inflammation. Thus, selective COX2 inhibition is a desirable characteristic of new anti-inflammatory drugs. The method of the invention can provide a route from identification of potential drug targets to validating these targets (for example, COX1 and COX2) as playing a role in disease (pain and inflammation) to an examination of the phenotype for the inhibition of one or both target isozymes without human suffering. Importantly, this information can be collected in vivo.

As an alternative strategy, the method of the invention can be used to define the phenotype of "genes of unknown function" obtained from various human genome sequencing projects or to assess the phenotype resulting, from inhibition of one isozyme subtype or one member of a family of related protein targets.

## **6.8. Biological Chips**

### **6.8.1. General Considerations**

In one aspect the present invention provides arrays of oligonucleotide probes immobilized in microfabricated patterns on silica chips for analyzing molecular interactions of biological interest.

The invention therefore relates to diverse fields impacted by the nature of molecular interaction, including chemistry, biology, medicine, and medical diagnostics.

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid). In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid.

See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose. See U.S. Patent Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference. Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. New methods and reagents are required to realize this promise, and the present invention helps meet that need.

### **6.8.2. General Strategies**

The invention provides several strategies employing immobilized arrays of probes for comparing a reference sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of, e.g.,

mutations. In a first embodiment, the invention provides a tiling strategy employing an array of immobilized oligonucleotide probes comprising at least two sets of probes. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. A second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. The probes in the first probe set have at least two interrogation positions corresponding to two contiguous nucleotides in the reference sequence. One interrogation position corresponds to one of the contiguous nucleotides, and the other interrogation position to the other.

In a second embodiment, the invention provides a tiling strategy employing an array comprising four probe sets. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. Second, third and fourth probe sets each comprise a corresponding probe for each probe in the first probe set.

The probes in the second, third and fourth probe sets are identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets. The first probe set often has at least 100 interrogation positions corresponding to 100 contiguous nucleotides in the reference sequence. Sometimes the first probe set has an interrogation position corresponding to every nucleotide in the reference sequence. The segment of complementarity within the probe set is usually about 9-21 nucleotides. Although probes may contain leading or

trailing sequences in addition to the 9-21 sequences, many probes consist exclusively of a 9-21 segment of complementarity.

In a third embodiment, the invention provides immobilized arrays of probes tiled for multiple reference sequences. one such array comprises at least one pair of first and second probe groups, each group comprising first and second sets of probes as defined in the first embodiment. Each probe in the first probe set from the first group is exactly complementary to a subsequence of a first reference sequence, and each probe in the first probe set from the second group is exactly complementary to a subsequence of a second reference sequence.

Thus, the first group of probes are tiled with respect to a first reference sequence and the second group of probes with respect to a second reference sequence. Each group of probes can also include third and fourth sets of probes as defined in the second embodiment. In some arrays of this type, the second reference sequence is a mutated form of the first reference sequence.

In a fourth embodiment, the invention provides arrays for block tiling. Block tiling is a species of the general tiling strategies described above. The usual unit of a block tiling array is a group of probes comprising a wildtype probe, a first set of three mutant probes and a second set of three mutant probes. The wildtype probe comprises a segment of at least three nucleotides exactly complementary to a subsequence of a reference sequence. The segment has at least first and second interrogation positions corresponding to first and second nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the first interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to a sequence comprising the wildtype probes or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the second interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe.



In a fifth embodiment, the invention provides methods of comparing a target sequence with a reference sequence using arrays of immobilized pooled probes. The arrays employed in these methods represent a further species of the general tiling arrays noted above. In these methods, variants of a reference sequence differing from the reference sequence in at least one nucleotide are identified and each is assigned a designation. An array of pooled probes is provided, with each pool occupying a separate cell of the array. Each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular designation.

The array is then contacted with a target sequence comprising a variant of the reference sequence. The relative hybridization intensities of the pools in the array to the target sequence are determined. The identity of the target sequence is deduced from the pattern of hybridization intensities. Often, each variant is assigned a designation having at least one digit and at least one value for the digit. In this case, each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular value in a particular digit. When variants are assigned successive numbers in a numbering system of base  $m$  having  $n$  digits,  $n \times (m-1)$  pooled probes are used to assign each variant a designation.

In a sixth embodiment, the invention provides a pooled probe for trellis tiling, a further species of the general tiling strategy. In trellis tiling, the identity of a nucleotide in a target sequence is determined from a comparison of hybridization intensities of three pooled trellis probes. A pooled trellis probe comprises a segment exactly complementary to a subsequence of a reference sequence except at a first interrogation position occupied by a pooled nucleotide  $N$ , a second interrogation position occupied by a pooled nucleotide selected from the group of three consisting of (1)  $M$  or  $K$ , (2)  $R$  or  $Y$  and (3)  $S$  or  $W$ , and a third interrogation position occupied by a second pooled nucleotide selected from the group. The pooled nucleotide occupying the second interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the second pooled probe and reference sequence are maximally aligned, and the pooled nucleotide occupying the third interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the third

pooled probe and the reference sequence are maximally aligned. Standard IUPAC nomenclature is used for describing pooled nucleotides.

In trellis tiling, an array comprises at least first, second and third cells, respectively occupied by first, second and third pooled probes, each according to the generic description above. However, the segment of complementarity, location of interrogation positions, and selection of pooled nucleotide at each interrogation position may or may not differ between the three pooled probes subject to the following constraint. One of the three interrogation positions in each of the three pooled probes must align with the same corresponding nucleotide in the reference sequence.

This interrogation position must be occupied by a N in one of the pooled probes, and a different pooled nucleotide in each of the other two pooled probes.

In a seventh embodiment, the invention provides arrays for bridge tiling. Bridge tiling is a species of the general tiling strategies noted above, in which probes from the first probe set contain more than one segment of complementarity.

In bridge tiling, a nucleotide in a reference sequence is usually determined from a comparison of four probes. A first probe comprises at least first and second segments, each of at least three nucleotides and each exactly complementary to first and second subsequences of a reference sequences. The segments including at least one interrogation position corresponding to a nucleotide in the reference sequence.

Either (1) the first and second subsequences are noncontiguous in the reference sequence, or (2) the first and second subsequences are contiguous and the first and second segments are inverted relative to the first and second subsequences.

The arrays further comprises second, third and fourth probes, which are identical to a sequence comprising the first probe or a subsequence thereof comprising at least three nucleotides from each of the first and second segments, except in the at least one interrogation position, which differs in each of the probes. In a species of bridge tiling,

referred to as deletion tiling, the first and second subsequences are separated by one or two nucleotides in the reference sequence.

In an eighth embodiment, the invention provides arrays of probes for multiplex tiling. Multiplex tiling is a strategy, in which the identity of two nucleotides in a target sequence is determined from a comparison of the hybridization intensities of four probes, each having two interrogation positions. Each of the probes comprising a segment of at least 7 nucleotides that is exactly complementary to a subsequence from a reference sequence, except that the segment may or may not be exactly complementary at two interrogation positions. The nucleotides occupying the interrogation positions are selected by the following rules: (1) the first interrogation position is occupied by a different nucleotide in each of the four probes, (2) the second interrogation position is occupied by a different nucleotide in each of the four probes, (3) in first and second probes, the segment is exactly complementary to the subsequence, except at no more than one of the interrogation positions, (4) in third and fourth probes, the segment is exactly complementary to the subsequence, except at both of the interrogation positions.

In a ninth embodiment, the invention provides arrays of immobilized probes including helper mutations. Helper mutations are useful for, e.g., preventing self-annealing of probes having inverted repeats. In this strategy, the identity of a nucleotide in a target sequence is usually determined from a comparison of four probes. A first probe comprises a segment of at least 7 nucleotides exactly complementary to a subsequence of a reference sequence except at one or two positions, the segment including an interrogation position not at the one or two positions. The one or two positions are occupied by helper mutations.

Second, third and fourth mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence thereof including the interrogation position and the one or two positions, except in the interrogation position, which is occupied by a different nucleotide in each of the four probes.

In a tenth embodiment, the invention provides arrays of probes comprising at least two probe sets, but lacking a probe set comprising probes that are perfectly matched to a reference sequence. Such arrays are usually employed in methods in which both reference

and target sequence are hybridized to the array. The first probe set comprising a plurality of probes, each probe comprising a segment exactly complementary to a subsequence of at least 3 nucleotides of a reference sequence except at an interrogation position. The second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the interrogation position, except that the interrogation position is occupied by a different nucleotide in each of the two corresponding probes and the complement to the reference sequence.

In an eleventh embodiment, the invention provides methods of comparing a target sequence with a reference sequence comprising a predetermined sequence of nucleotides using any of the arrays described above. The methods comprise hybridizing the target nucleic acid to an array and determining which probes, relative to one another, in the array bind specifically to the target nucleic acid. The relative specific binding of the probes indicates whether the target sequence is the same or different from the reference sequence. In some such methods, the target sequence has a substituted nucleotide relative to the reference sequence in at least one undetermined position, and the relative specific binding of the probes indicates the location of the position and the nucleotide occupying the position in the target sequence. In some methods, a second target nucleic acid is also hybridized to the array. The relative specific binding of the probes then indicates both whether the target sequence is the same or different from the reference sequence, and whether the second target sequence is the same or different from the reference sequence. In some methods, when the array comprises two groups of probes tiled for first and second reference sequences, respectively, the relative specific binding of probes in the first group indicates whether the target sequence is the same or different from the first reference sequence. The relative specific binding of probes in the second group indicates whether the target sequence is the same or different from the second reference sequence. Such methods are particularly useful for analyzing heterologous alleles of a gene. Some methods entail hybridizing both a reference sequence and a target sequence to any of the arrays of probes described above. Comparison of the relative specific binding of the probes to the reference and target sequences indicates whether the target sequence is the same or different from the reference sequence.

In a twelfth embodiment, the invention provides arrays of immobilized probes in which the probes are designed to tile a reference sequence from a human immunodeficiency virus.

Reference sequences from either the reverse transcriptase gene or protease gene of HIV are of particular interest. Some chips further comprise arrays of probes tiling a reference sequence from a 16S RNA or DNA encoding the 16S RNA from a pathogenic microorganism. The invention further provides methods of using such arrays in analyzing a HIV target sequence. The methods are particularly useful where the target sequence has a substituted nucleotide relative to the reference sequence in at least one position, the substitution conferring resistance to a drug use in treating a patient infected with a HIV virus. The methods reveal the existence of the substituted nucleotide. The methods are also particularly useful for analyzing a mixture of undetermined proportions of first and second target sequences from different HIV variants. The relative specific binding of probes indicates the proportions of the first and second target sequences.

In a thirteenth embodiment, the invention provides arrays of probes tiled based on reference sequence from a CFTR gene. A preferred array comprises at least a group of probes comprising a wildtype probe, and five sets of three mutant probes. The wildtype probe is exactly complementary to a subsequence of a reference sequence from a cystic fibrosis gene, the segment having at least five interrogation positions corresponding to five contiguous nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to the wildtype probe, except in a first of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to the wildtype probe, except in a second of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the third set of three mutant probes are each identical to the wildtype probe, except in a third of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fourth set of three mutant probes are each identical to the wildtype probe, except in a fourth of the five interrogation positions, which is occupied by

a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fifth set of three mutant probes are each identical to the wildtype probe, except in a fifth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. Preferably, a chip comprises two such groups of probes. The first group comprises a wildtype probe exactly complementary to a first reference sequence, and the second group comprises a wildtype probe exactly complementary to a second reference sequence that is a mutated form of the first reference sequence.

The invention further provides methods of using the arrays of the invention for analyzing target sequences from a CFTR gene. The methods are capable of simultaneously analyzing first and second target sequences representing heterozygous alleles of a CFTR gene.

In a fourteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a p53 gene, an hMLH1 gene and/or an MSH2 gene. The invention further provides methods of using the arrays described above to analyze these genes. The methods are useful, e.g., for diagnosing patients susceptible to developing cancer.

In a fifteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a mitochondrial genome. The reference sequence may comprise part or all of the D-loop region, or all, or substantially all, of the mitochondrial genome. The invention further provides method of using the arrays described above to analyze target sequences from a mitochondrial genome. The methods are useful for identifying mutations associated with disease, and for forensic, epidemiological and evolutionary studies.

### **6.8.3. Specific Strategies**

The invention provides a number of strategies for comparing a polynucleotide of known sequence (a reference sequence) with variants of that sequence (target sequences).

The comparison can be performed at the level of entire genomes, chromosomes, genes, exons or introns, or can focus on individual mutant sites and immediately adjacent bases. The strategies allow detection of variations, such as mutations or polymorphisms, in the

target sequence irrespective whether a particular variant has previously been characterized. The strategies both define the nature of a variant and identify its location in a target sequence.

The strategies employ arrays of oligonucleotide probes immobilized to a solid support. Target sequences are analyzed by determining the extent of hybridization at particular probes in the array. The strategy in selection of probes facilitates distinction between perfectly matched probes and probes showing single-base or other degrees of mismatches.

The strategy usually entails sampling each nucleotide of interest in a target sequence several times, thereby achieving a high degree of confidence in its identity. This level of confidence is further increased by sampling of adjacent nucleotides in the target sequence to nucleotides of interest.

The number of probes on the chip can be quite large (e.g.,  $10^5$ - $10^6$ ). However, usually only a small proportion of the total number of probes of a given length are represented.

Some advantage of the use of only a small proportion of all possible probes of a given length include: (i) each position in the array is highly informative, whether or not hybridization occurs; (ii) nonspecific hybridization is minimized; (iii) it is straightforward to correlate hybridization differences with sequence differences, particularly with reference to the hybridization pattern of a known standard; and (iv) the ability to address each probe independently during synthesis, using high resolution photolithography, allows the array to be designed and optimized for any sequence. For example the length of any probe can be varied independently of the others.

The present tiling strategies result in sequencing and comparison methods suitable for routine large-scale practice with a high degree of confidence in the sequence output.

### 6.8.3.1. General Tiling Strategies

#### 6.8.3.1.1. Selection of Reference Sequence

The chips are designed to contain probes exhibiting complementarity to one or more selected reference sequence whose sequence is known. The chips are used to read a target sequence comprising either the reference sequence itself or variants of that sequence. Target sequences may differ from the reference sequence at one or more positions but show a high overall degree of sequence identity with the reference sequence (e.g., at least 75, 90, 95, 99, 99.9 or 99-99%). Any polynucleotide of known sequence can be selected as a reference sequence. Reference sequences of interest include sequences known to include mutations or polymorphisms associated with phenotypic changes having clinical significance in human patients. For example, the CFTR gene and P53 gene in humans have been identified as the location of several mutations resulting in cystic fibrosis or cancer respectively. Other reference sequences of interest include those that serve to identify pathogenic microorganisms and/or are the site of mutations by which such microorganisms acquire drug resistance (e.g., the HIV reverse transcriptase gene). Other reference sequences of interest include regions where polymorphic variations are known to occur (e.g., the D-loop region of mitochondrial DNA). These reference sequences have utility for, e.g., forensic or epidemiological studies. Other reference sequences of interest include p34 (related to p53), p65 (implicated in breast, prostate and liver cancer), and DNA segments encoding cytochromes P450 (see Meyer et al., *Pharmac. Ther.* 46, 349-355 (1990)). Other reference sequences of interest include those from the genome of pathogenic viruses (e.g., hepatitis J, B, or Q, herpes virus (e.g., VZV, HSV-1, HAV-6, HSV-II, and CMV, Epstein Barr virus), adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, cornovirus, respiratory syncytial virus, mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, molluscum virus, poliovirus, rabies virus, JC virus and arboviral encephalitis virus. Other reference sequences of interest are from genomes or episomes of pathogenic bacteria, particularly regions that confer drug resistance or allow phylogenic characterization of the host (e.g., 16S rRNA or corresponding DNA). For example, such bacteria include chlamydia, rickettsial bacteria, mycobacteria, staphylococci, treptocci, pneumonococci, meningococci and conococci, klebsiella,



proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lyme disease bacteria. Other reference sequences of interest include those in which mutations result in the following autosomal recessive disorders: sickle cell anemia, beta-thalassemia, phenylketonuria, galactosemia, Wilson's disease, hemochromatosis, severe combined immunodeficiency, alpha-1-antitrypsin deficiency, albinism, alkaptonuria, lysosomal storage diseases and Ehlers-Danlos syndrome. Other reference sequences of interest include those in which mutations result in X-linked recessive disorders: hemophilia, glucose-6-phosphate dehydrogenase, agammaglobulinemia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease and fragile X- syndrome. Other reference sequences of interest includes those in which mutations result in the following autosomal dominant disorders: familial hypercholesterolemia, polycystic kidney disease, Huntingdon's disease, hereditary spherocytosis, Marfan's syndrome, von Willebrand's disease, neurofibromatosis, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, myotonic dystrophy, muscular dystrophy, osteogenesis imperfecta, acute intermittent porphyria, and von Hippel- Lindau disease.

The length of a reference sequence can vary widely from a full-length genome, to an individual chromosome, episome, gene, component of a gene, such as an exon, intron or regulatory sequences, to a few nucleotides. A reference sequence of between about 2, 5, 10, 20, 50, 100, 500, 1000, 5,000 or 10,000, 20,000 or 100,000 nucleotides is common.

Sometimes only particular regions of a sequence (e.g., exons of a gene) are of interest. In such situations, the particular regions can be considered as separate reference sequences or can be considered as components of a single reference sequence, as matter of arbitrary choice.

A reference sequence can be any naturally occurring, mutant, consensus or purely hypothetical sequence of nucleotides, RNA or DNA. For example, sequences can be obtained from computer data bases, publications or can be determined or conceived de novo. Usually, a reference sequence is selected to show a high degree of sequence identity to envisaged target sequences. Often, particularly, where a significant degree of

divergence is anticipated between target sequences, more than one reference sequence is selected. Combinations of wildtype and mutant reference sequences are employed in several applications of the tiling strategy.

#### **6.8.3.1.2. Chip Design**

##### **6.8.3.1.2.1. Basic Tiling Strategy**

The basic tiling strategy provides an array of immobilized probes for analysis of target sequences showing a high degree of sequence identity to one or more selected reference sequences. The strategy is first illustrated for an array that is subdivided into four probe sets, although it will be apparent that in some situations, satisfactory results are obtained from only two probe sets. A first probe set comprises a plurality of probes exhibiting perfect complementarity with a selected reference sequence. The perfect complementarity usually exists throughout the length of the probe. However, probes having a segment or segments of perfect complementarity that is/are flanked by leading or trailing sequences lacking complementarity to the reference sequence can also be used. Within a segment of complementarity, each probe in the first probe set has at least one interrogation position that corresponds to a nucleotide in the reference sequence. That is, the interrogation position is aligned with the corresponding nucleotide in the reference sequence, when the probe and reference sequence are aligned to maximize complementarity between the two. If a probe has more than one interrogation position, each corresponds with a respective nucleotide in the reference sequence. The identity of an interrogation position and corresponding nucleotide in a particular probe in the first probe set cannot be determined simply by inspection of the probe in the first set. As will become apparent, an interrogation position and corresponding nucleotide is defined by the comparative structures of probes in the first probe set and corresponding probes from additional probe sets.

In principle, a probe could have an interrogation position at each position in the segment complementary to the reference sequence. Sometimes, interrogation positions provide more accurate data when located away from the ends of a segment of complementarity. Thus, typically a probe having a segment of complementarity of length  $x$  does not contain

more than  $x-2$  interrogation positions. Since probes are typically 9-21 nucleotides, and usually all of a probe is complementary, a probe typically has 1-19 interrogation positions. Often the probes contain a single interrogation position, at or near the center of probe.

For each probe in the first set, there are, for purposes of the present illustration, three corresponding probes from three additional probe sets. Thus, there are four probes corresponding to each nucleotide of interest in the reference sequence. Each of the four corresponding probes has an interrogation position aligned with that nucleotide of interest. Usually, the probes from the three additional probe sets are identical to the corresponding probe from the first probe set with one exception. The exception is that at least one (and often only one) interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, is occupied by a different nucleotide in the four probe sets. For example, for an A nucleotide in the reference sequence, the corresponding probe from the first probe set has its interrogation position occupied by a T, and the corresponding probes from the additional three probe sets have their respective interrogation positions occupied by A, C, or G, a different nucleotide in each probe. Of course, if a probe from the first probe set comprises trailing or flanking sequences lacking complementarity to the reference sequences, these sequences need not be present in corresponding probes from the three additional sets. Likewise corresponding probes from the three additional sets can contain leading or trailing sequences outside the segment of complementarity that are not present in the corresponding probe from the first probe set. Occasionally, the probes from the additional three probe set are identical (with the exception of interrogation position(s)) to a contiguous subsequence of the full complementary segment of the corresponding probe from the first probe set. In this case, the subsequence includes the interrogation position and usually differs from the full-length probe only in the omission of one or both terminal nucleotides from the termini of a segment of complementarity.

That is, if a probe from the first probe set has a segment of complementarity of length  $n$ , corresponding probes from the other sets will usually include a subsequence of the segment of at least length  $n-2$ . Thus, the subsequence is usually at least 3, 4, 7, 9, 15, 21, or 25 nucleotides long, most typically, in the range of 9-21 nucleotides. The subsequence should be sufficiently long to allow a probe to hybridize detectably more strongly to a

variant of the reference sequence mutated at the interrogation position than to the reference sequence.

The probes can be oligodeoxyribonucleotides or oligoribonucleotides, or any modified forms of these polymers that are capable of hybridizing with a target nucleic sequence by complementary base-pairing. Complementary base pairing means sequence-specific base pairing which includes e.g., Watson-Crick base pairing as well as other forms of base pairing such as Hoogsteen base pairing. Modified forms include 2'-O-methyl oligoribonucleotides and so-called PNAs, in which oligodeoxyribonucleotides are linked via peptide bonds rather than phosphodiester bonds. The probes can be attached by any linkage to a support (e.g., 3', 5' or via the base). 3' attachment is more usual as this orientation is compatible with the preferred chemistry for solid phase synthesis of oligonucleotides.

The number of probes in the first probe set (and as a consequence the number of probes in additional probe sets) depends on the length of the reference sequence, the number of nucleotides of interest in the reference sequence and the number of interrogation positions per probe. In general, each nucleotide of interest in the reference sequence requires the same interrogation position in the four sets of probes.

Consider, as an example, a reference sequence of 100 nucleotides, 50 of which are of interest, and probes each having a single interrogation position. In this situation, the first probe set requires fifty probes, each having one interrogation position corresponding to a nucleotide of interest in the reference sequence. The second, third and fourth probe sets each have a corresponding probe for each probe in the first probe set, and so each also contains a total of fifty probes. The identity of each nucleotide of interest in the reference sequence is determined by comparing the relative hybridization signals at four probes having interrogation positions corresponding to that nucleotide from the four probe sets.

In some reference sequences, every nucleotide is of interest. In other reference sequences, only certain portions in which variants (e.g., mutations or polymorphisms) are concentrated are of interest. In other reference sequences, only particular mutations or polymorphisms and immediately adjacent nucleotides are of interest. Usually, the first

probe set has interrogation positions selected to correspond to at least a nucleotide (e.g., representing a point mutation) and one immediately adjacent nucleotide. Usually, the probes in the first set have interrogation positions corresponding to at least 3, 10, 50, 100, 1000, or 20,000 contiguous nucleotides. The probes usually have interrogation positions corresponding to at least 5, 10, 30, 50, 75, 90, 99 or sometimes 100% of the nucleotides in a reference sequence.

Frequently, the probes in the first probe set completely span the reference sequence and overlap with one another relative to the reference sequence. For example, in one common arrangement each probe in the first probe set differs from another probe in that set by the omission of a 3' base complementary to the reference sequence and the acquisition of a 5' base complementary to the reference sequence.

For conceptual simplicity, the probes in a set are usually arranged in order of the sequence in a lane across the chip. A lane contains a series of overlapping probes, which represent or tile across, the selected reference sequence. The components of the four sets of probes are usually laid down in four parallel lanes, collectively constituting a row in the horizontal direction and a series of 4-member columns in the vertical direction. Corresponding probes from the four probe sets (i.e., complementary to the same subsequence of the reference sequence) occupy a column.

Each probe in a lane usually differs from its predecessor in the lane by the omission of a base at one end and the inclusion of additional base at the other end. However, this orderly progression of probes can be interrupted by the inclusion of control probes or omission of probes in certain columns of the array. Such columns serve as controls to orient the chip, or gauge the background, which can include target sequence nonspecifically bound to the chip.

The probes sets are usually laid down in lanes such that all probes having an interrogation position occupied by an A form an A-lane, all probes having an interrogation position occupied by a C form a C-lane, all probes having an interrogation position occupied by a G form a G-lane, and all probes having an interrogation position occupied by a T (or U) form a T lane (or a U lane). Note that in this arrangement there is not a unique

correspondence between probe sets and lanes. Thus, the probe from the first probe set is laid down in the A-lane, C-lane, A-lane, A-lane and T-lane for the five columns. The interrogation position on a column of probes corresponds to the position in the target sequence whose identity is determined from analysis of hybridization to the probes in that column. The interrogation position can be anywhere in a probe but is usually at or near the central position of the probe to maximize differential hybridization signals between a perfect match and a single-base mismatch.

For example, for an 11 mer probe, the central position is the sixth nucleotide.

Although the array of probes is usually laid down in rows and columns as described above, such a physical arrangement of probes on the chip is not essential. Provided that the spatial location of each probe in an array is known, the data from the probes can be collected and processed to yield the sequence of a target irrespective of the physical arrangement of the probes on a chip. In processing the data, the hybridization signals from the respective probes can be reassorted into any conceptual array desired for subsequent data reduction whatever the physical arrangement of probes on the chip.

A range of lengths of probes can be employed in the chips. As noted above, a probe may consist exclusively of a complementary segments, or may have one or more complementary segments juxtaposed by flanking, trailing and/or intervening segments. In the latter situation, the total length of complementary segment(s) is more important than the length of the probe. In functional terms, the complementarity segment(s) of the first probe sets should be sufficiently long to allow the probe to hybridize detectably more strongly to a reference sequence compared with a variant of the reference including a single base mutation at the nucleotide corresponding to the interrogation position of the probe.

Similarly, the complementarity segment(s) in corresponding probes from additional probe sets should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence having a single nucleotide substitution at the interrogation position relative to the reference sequence. A probe usually has a single complementary segment having a length of at least 3 nucleotides, and more usually at least

5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 or bases exhibiting perfect complementarity (other than possibly at the interrogation position(s) depending on the probe set) to the reference sequence. In bridging strategies, where more than one segment of complementarity is present, each segment provides at least three complementary nucleotides to the reference sequence and the combined segments provide at least two segments of three or a total of six complementary nucleotides. As in the other strategies, the combined length of complementary segments is typically from 6-30 nucleotides, and preferably from about 9-21 nucleotides. The two segments are often approximately the same length. Often, the probes (or segment of complementarity within probes) have an odd number of bases, so that an interrogation position can occur in the exact center of the probe.

In some chips, all probes are the same length. Other chips employ different groups of probe sets, in which case the probes are of the same size within a group, but differ between different groups. For example, some chips have one group comprising four sets of probes as described above in which all the probes are 11 mers, together with a second group comprising four sets of probes in which all of the probes are 13 mers. Of course, additional groups of probes can be added.

Thus, some chips contain, e.g., four groups of probes having sizes of 11 mers, 13 mers, 15 mers and 17 mers. Other chips have different size probes within the same group of four probe sets. In these chips, the probes in the first set can vary in length independently of each other. Probes in the other sets are usually the same length as the probe occupying the same column from the first set. However, occasionally different lengths of probes can be included at the same column position in the four lanes. The different length probes are included to equalize hybridization signals from probes irrespective of whether A-T or C-G bonds are formed at the interrogation position.

The length of probe can be important in distinguishing between a perfectly matched probe and probes showing a single- base mismatch with the target sequence. The discrimination is usually greater for short probes. Shorter probes are usually also less susceptible to formation of secondary structures.

However, the absolute amount of target sequence bound, and hence the signal, is greater for larger probes. The probe length representing the optimum compromise between these competing considerations may vary depending on inter alia the GC content of a particular region of the target DNA sequence, secondary structure, synthesis efficiency and cross-hybridization. In some regions of the target, depending on hybridization conditions, short probes (e.g., 11 mers) may provide information that is inaccessible from longer probes (e.g., 19 mers) and vice versa. Maximum sequence information can be read by including several groups of different sized probes on the chip as noted above. However, for many regions of the target sequence, such a strategy provides redundant information in that the same sequence is read multiple times from the different groups of probes. Equivalent information can be obtained from a single group of different sized probes in which the sizes are selected to maximize readable sequence at particular regions of the target sequence. The strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences.

The invention provides an optimization block which allows systematic variation of probe length and interrogation position to optimize the selection of probes for analyzing a particular nucleotide in a reference sequence. The block comprises alternating columns of probes complementary to the wildtype target and probes complementary to a specific mutation. The interrogation position is varied between columns and probe length is varied down a column.

Hybridization of the chip to the reference sequence or the mutant form of the reference sequence identifies the probe length and interrogation position providing the greatest differential hybridization signal.

The probes are designed to be complementary to either strand of the reference sequence (e.g., coding or non-coding). some chips contain separate groups of probes, one complementary to the coding strand, the other complementary to the noncoding strand. Independent analysis of coding and noncoding strands provides largely redundant information.



However, the regions of ambiguity in reading the coding strand are not always the same as those in reading the noncoding strand. Thus, combination of the information from coding and noncoding strands increases the overall accuracy of sequencing.

Some chips contain additional probes or groups of probes designed to be complementary to a second reference sequence.

The second reference sequence is often a subsequence of the first reference sequence bearing one or more commonly occurring mutations or interstrain variations. The second group of probes is designed by the same principles as described above except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group is particularly useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases). Of course, the same principle can be extended to provide chips containing groups of probes for any number of reference sequences. Alternatively, the chips may contain additional probe(s) that do not form part of a tiled array as noted above, but rather serves as probe(s) for a conventional reverse dot blot. For example, the presence of mutation can be detected from binding of a target sequence to a single oligomeric probe harboring the mutation. Preferably, an additional probe containing the equivalent region of the wildtype sequence is included as a control.

The chips are read by comparing the intensities of labelled target bound to the probes in an array.

Specifically, a comparison is performed between each lane of probes (e.g., A, C, G and T lanes) at each columnar position (physical or conceptual). For a particular columnar position, the lane showing the greatest hybridization signal is called as the nucleotide present at the position in the target sequence corresponding to the interrogation position in the probes. The corresponding position in the target sequence is that aligned with the interrogation position in corresponding probes when the probes and target are aligned to maximize complementarity. Of the four probes in a column, only one can exhibit a perfect match to the target sequence whereas the others usually exhibit at least a one base pair

mismatch. The probe exhibiting a perfect match usually produces a substantially greater hybridization signal than the other three probes in the column and is thereby easily identified. However, in some regions of the target sequence, the distinction between a perfect match and a one-base mismatch is less clear. Thus, a call ratio is established to define the ratio of signal from the best hybridizing probes to the second best hybridizing probe that must be exceeded for a particular target position to be read from the probes. A high call ratio ensures that few if any errors are made in calling target nucleotides, but can result in some nucleotides being scored as ambiguous, which could in fact be accurately read.

A lower call ratio results in fewer ambiguous calls, but can result in more erroneous calls. It has been found that at a call ratio of 1.2 virtually all calls are accurate. However, a small but significant number of bases (e.g., up to about %) may have to be scored as ambiguous.

Although small regions of the target sequence can sometimes be ambiguous, these regions usually occur at the same or similar segments in different target sequences. Thus, for precharacterized mutations, it is known in advance whether that mutation is likely to occur within a region of unambiguously determinable sequence.

An array of probes is most useful for analyzing the reference sequence from which the probes were designed and variants of that sequence exhibiting substantial sequence similarity with the reference sequence (e.g., several single- base mutants spaced over the reference sequence). When an array is used to analyze the exact reference sequence from which it was designed, one probe exhibits a perfect match to the reference sequence, and the other three probes in the same column exhibit single-base mismatches. Thus, discrimination between hybridization signals is usually high and accurate sequence is obtained. High accuracy is also obtained when an array is used for analyzing a target sequence comprising a variant of the reference sequence that has a single mutation relative to the reference sequence, or several widely spaced mutations relative to the reference sequence. At different mutant loci, one probe exhibits a perfect match to the target, and the other three probes occupying the same column exhibit single-base mismatches, the difference (with respect to analysis of the reference sequence) being the lane in which the perfect match occurs.

For target sequences showing a high degree of divergence from the reference strain or incorporating several closely spaced mutations from the reference strain, a single group of probes (i.e., designed with respect to a single reference sequence) will not always provide accurate sequence for the highly variant region of this sequence. At some particular columnar positions, it may be that no single probe exhibits perfect complementarity to the target and that any comparison must be based on different degrees of mismatch between the four probes. Such a comparison does not always allow the target nucleotide corresponding to that columnar position to be called. Deletions in target sequences can be detected by loss of signal from probes having interrogation positions encompassed by the deletion. However, signal may also be lost from probes having interrogation positions closely proximal to the deletion resulting in some regions of the target sequence that cannot be read. Target sequence bearing insertions will also exhibit short regions including and proximal to the insertion that usually cannot be read.

The presence of short regions of difficult-to-read target because of closely spaced mutations, insertions or deletion, does not prevent determination of the remaining sequence of the target as different regions of a target sequence are determined independently. Moreover, such ambiguities as might result from analysis of diverse variants with a single group of probes can be avoided by including multiple groups of probe sets on a chip. For example, one group of probes can be designed based on a full-length reference sequence, and the other groups on subsequences of the reference sequence incorporating frequently occurring mutations or strain variations.

A particular advantage of the present sequencing strategy over conventional sequencing methods is the capacity simultaneously to detect and quantify proportions of multiple target sequences. Such capacity is valuable, e.g., for diagnosis of patients who are heterozygous with respect to a gene or who are infected with a virus, such as HIV, which is usually present in several polymorphic forms. Such capacity is also useful in analyzing targets from biopsies of tumor cells and surrounding tissues. The presence of multiple target sequences is detected from the relative signals of the four probes at the array columns corresponding to the target nucleotides at which diversity occurs. The relative signals at the four probes for the mixture under test are compared with the corresponding signals from a homogeneous reference sequence. An increase in a signal from a probe that

is mismatched with respect to the reference sequence, and a corresponding decrease in the signal from the probe which is matched with the reference sequence signal the presence of a mutant strain in the mixture. The extent in shift in hybridization signals of the probes is related to the proportion of a target sequence in the mixture. Shifts in relative hybridization signals can be quantitatively related to proportions of reference and mutant sequence by prior calibration of the chip with seeded mixtures of the mutant and reference sequences. By this means, a chip can be used to detect variant or mutant strains constituting as little as 1, 5, 20, or 25 % of a mixture of stains.

Similar principles allow the simultaneous analysis of multiple target sequences even when none is identical to the reference sequence. For example, with a mixture of two target sequences bearing first and second mutations, there would be a variation in the hybridization patterns of probes having interrogation positions corresponding to the first and second mutations relative to the hybridization pattern with the reference sequence. At each position, one of the probes having a mismatched interrogation position relative to the reference sequence would show an increase in hybridization signal, and the probe having a matched interrogation position relative to the reference sequence would show a decrease in hybridization signal. Analysis of the hybridization pattern of the mixture of mutant target sequences, preferably in comparison with the hybridization pattern of the reference sequence, indicates the presence of two mutant target sequences, the position and nature of the mutation in each strain, and the relative proportions of each strain.

In a variation of the above method, the different components in a mixture of target sequences are differentially labelled before being applied to the array. For example, a variety of fluorescent labels emitting at different wavelength are available. The use of differential labels allows independent analysis of different targets bound simultaneously to the array. For example, the methods permit comparison of target sequences obtained from a patient at different stages of a disease.

#### 6.8.3.1.2.2. Omission of Probes

The general strategy outlined above employs four probes to read each nucleotide of interest in a target sequence. One probe (from the first probe set) shows a perfect match to the reference sequence and the other three probes (from the second, third and fourth probe sets) exhibit a mismatch with the reference sequence and a perfect match with a target sequence bearing a mutation at the nucleotide of interest.

The provision of three probes from the second, third and fourth probe sets allows detection of each of the three possible nucleotide substitutions of any nucleotide of interest.

However, in some reference sequences or regions of reference sequences, it is known in advance that only certain mutations are likely to occur. Thus, for example, at one site it might be known that an A nucleotide in the reference sequence may exist as a T mutant in some target sequences but is unlikely to exist as a C or G mutant. Accordingly, for analysis of this region of the reference sequence, one might include only the first and second probe sets, the first probe set exhibiting perfect complementarity to the reference sequence, and the second probe set having an interrogation position occupied by an invariant A residue (for detecting the T mutant). In other situations, one might include the first, second and third probes sets (but not the fourth) for detection of a wildtype nucleotide in the reference sequence and two mutant variants thereof in target sequences. In some chips, probes that would detect silent mutations (i.e., not affecting amino acid sequence) are omitted.

In some chips, the probes from the first probe set are omitted corresponding to some or all positions of the reference sequences. Such chips comprise at least two probe sets. The first probe set has a plurality of probes. Each probe comprises a segment exactly complementary to a subsequence of a reference sequence except in at least one interrogation position. A second probe set has a corresponding probe for each probe in the first probe set.

The corresponding probe in the second probe set is identical to a sequence comprising the corresponding probe from the first probe set or a subsequence thereof that includes the at least one (and usually only one) interrogation position except that the at least one

interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. A third probe set, if present, also comprises a corresponding probe for each probe in the first probe set except at the at least one interrogation position, which differs in the corresponding probes from the three sets. Omission of probes having a segment exhibiting perfect complementarity to the reference sequence results in loss of control information, i.e., the detection of nucleotides in a target sequence that are the same as those in a reference sequence. However, similar information can be obtained by hybridizing a chip lacking probes from the first probe set to both target and reference sequences. The hybridization can be performed sequentially, or concurrently, if the target and reference are differentially labelled. In this situation, the presence of a mutation is detected by a shift in the background hybridization intensity of the reference sequence to a perfectly matched hybridization signal of the target sequence, rather than by a comparison of the hybridization intensities of probes from the first set with corresponding probes from the second, third and fourth sets.

#### **6.8.3.1.2.3. Wildtype Probe Lane**

When the chips comprise four probe sets, as discussed supra, and the probe sets are laid down in four lanes, an A-lane, a C-lane, a G-lane and a T or U-lane, the probe having a segment exhibiting perfect complementarity to a reference sequence varies between the four lanes from one column to another. This does not present any significant difficulty in computer analysis of the data from the chip. However, visual inspection of the hybridization pattern of the chip is sometimes facilitated by provision of an extra lane of probes, in which each probe has a segment exhibiting perfect complementarity to the reference sequence. This segment is identical to a segment from one of the probes in the other four lanes (which lane depending on the column position). The extra lane of probes (designated the wildtype lane) hybridizes to a target sequence at all nucleotide positions except those in which deviations from the reference sequence occurs. The hybridization pattern of the wildtype lane thereby provides a simple visual indication of mutations.

#### 6.8.3.1.2.4. Deletion, Insertion and Multiple-Mutation Probes

Some chips provide an additional probe set specifically designed for analyzing deletion mutations. The additional probe set comprises a probe corresponding to each probe in the first probe set as described above. However, a probe from the additional probe set differs from the corresponding probe in the first probe set in that the nucleotide occupying the interrogation position is deleted in the probe from the additional probe set. Optionally, the probe from the additional probe set bears an additional nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set will hybridize more strongly than the corresponding probe from the first probe set to a target sequence having a single base deletion at the nucleotide corresponding to the interrogation position. Additional probe sets are provided in which not only the interrogation position, but also an adjacent nucleotide is detected.

Similarly, other chips provide additional probe sets for analyzing insertions. For example, one additional probe set has a probe corresponding to each probe in the first probe set as described above. However, the probe in the additional probe set has an extra T nucleotide inserted adjacent to the interrogation position. Optionally, the probe has one fewer nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set hybridizes more strongly than the corresponding probe from the first probe set to a target sequence having an A nucleotide inserted in a position adjacent to that corresponding to the interrogation position.

Similar additional probe sets are constructed having C, G or T/U nucleotides inserted adjacent to the interrogation position. Usually, four such probe sets, one for each nucleotide, are used in combination.

Other chips provide additional probes (multiple-mutation probes) for analyzing target sequences having multiple closely spaced mutations. A multiple-mutation probe is usually identical to a corresponding probe from the first set as described above, except in the base occupying the interrogation position, and except at one or more additional positions, corresponding to nucleotides in which substitution may occur in the reference sequence. The one or more additional positions in the multiple mutation probe are occupied by

nucleotides complementary to the nucleotides occupying corresponding positions in the reference sequence when the possible substitutions have occurred.

#### 6.8.3.1.2.5. Block Tiling

As noted in the discussion of the general tiling strategy, a probe in the first probe set sometimes has more than one interrogation position. In this situation, a probe in the first probe set is sometimes matched with multiple groups of at least one, and usually, three additional probe sets. Three additional probe sets are used to allow detection of the three possible nucleotide substitutions at any one position. If only certain types of substitution are likely to occur (e.g., transitions), only one or two additional probe sets are required (analogous to the use of probes in the basic tiling strategy). To illustrate for the situation where a group comprises three additional probe sets, a first such group comprises second, third and fourth probe sets, each of which has a probe corresponding to each probe in the first probe set. The corresponding probes from the second, third and fourth probe sets differ from the corresponding probe in the first set at a first of the interrogation positions. Thus, the relative hybridization signals from corresponding probes from the first, second, third and fourth probe sets indicate the identity of the nucleotide in a target sequence corresponding to the first interrogation position. A second group of three probe sets (designated fifth, sixth and seventh probe sets), each also have a probe corresponding to each probe in the first probe set. These corresponding probes differ from that in the first probe set at a second interrogation position. The relative hybridization signals from corresponding probes from the first, fifth, sixth, and seventh probe sets indicate the identity of the nucleotide in the target sequence corresponding to the second interrogation position. As noted above, the probes in the first probe set often have seven or more interrogation positions. If there are seven interrogation positions, there are seven groups of three additional probe sets, each group of three probe sets serving to identify the nucleotide corresponding to one of the seven interrogation positions.

Each block of probes allows short regions of a target sequence to be read. For example, for a block of probes having seven interrogation positions, seven nucleotides in the target sequence can be read. Of course, a chip can contain any number of blocks depending on how many nucleotides of the target are of interest. The hybridization signals for each



block can be analyzed independently of any other block. The block tiling strategy can also be combined with other tiling strategies, with different parts of the same reference sequence being tiled by different strategies.

The block tiling strategy offers two advantages over the basic strategy in which each probe in the first set has a single interrogation position. One advantage is that the same sequence information can be obtained from fewer probes. A second advantage is that each of the probes constituting a block (i.e., a probe from the first probe set and a corresponding probe from each of the other probe sets) can have identical 3' and 5' sequences, with the variation confined to a central segment containing the interrogation positions. The identity of 3' sequence between different probes simplifies the strategy for solid phase synthesis of the probes on the chip and results in more uniform deposition of the different probes on the chip, thereby in turn increasing the uniformity of signal to noise ratio for different regions of the chip. A third advantage is that greater signal uniformity is achieved within a block.

#### 6.8.3.1.2.6. Multiplex Tiling

In the block tiling strategy discussed above, the identity of a nucleotide in a target or reference sequence is determined by comparison of hybridization patterns of one probe having a segment showing a perfect match with that of other probes (usually three other probes) showing a single base mismatch. In multiplex tiling, the identity of at least two nucleotides in a reference or target sequence is determined by comparison of hybridization signal intensities of four probes, two of which have a segment showing perfect complementarity or a single base mismatch to the reference sequence, and two of which have a segment showing perfect complementarity or a double-base mismatch to a segment. The four probes whose hybridization patterns are to be compared each have a segment that is exactly complementary to a reference sequence except at two interrogation positions, in which the segment may or may not be complementary to the reference sequence. The interrogation positions correspond to the nucleotides in a reference or target sequence which are determined by the comparison of intensities. The nucleotides occupying the interrogation positions in the four probes are selected according to the following rule. The first interrogation position is occupied by a different nucleotide in each of the four probes.

The second interrogation position is also occupied by a different nucleotide in each of the four probes. In two of the four probes, designated the first and second probes, the segment is exactly complementary to the reference sequence except at not more than one of the two interrogation positions. In other words, one of the interrogation positions is occupied by a nucleotide that is complementary to the corresponding nucleotide from the reference sequence and the other interrogation position may or may not be so occupied. In the other two of the four probes, designated the third and fourth probes, the segment is exactly complementary to the reference sequence except that both interrogation positions are occupied by nucleotides which are noncomplementary to the respective corresponding nucleotides in the reference sequence.

There are number of ways of satisfying these conditions depending on whether the two nucleotides in the reference sequence corresponding to the two interrogation positions are the same or different. If these two nucleotides are different in the reference sequence (probability 3/4), the conditions are satisfied by each of the two interrogation positions being occupied by the same nucleotide in any given probe. For example, in the first probe, the two interrogation positions would both be A, in the second probe, both would be C, in the third probe, each would be G, and in the fourth probe each would be T or U. If the two nucleotides in the reference sequence corresponding to the two interrogation positions are different, the conditions noted above are satisfied by each of the interrogation positions in any one of the four probes being occupied by complementary nucleotides. For example, in the first probe, the interrogation positions could be occupied by A and T, in the second probe by C and G, in the third probe by G and C, and in the fourth probe, by T and A.

When the four probes are hybridized to a target that is the same as the reference sequence or differs from the reference sequence at one (but not both) of the interrogation positions, two of the four probes show a double-mismatch with the target and two probes show a single mismatch. The identity of probes showing these different degrees of mismatch can be determined from the different hybridization signals.

From the identity of the probes showing the different degrees of mismatch, the nucleotides occupying both of the interrogation positions in the target sequence can be deduced.

For ease of illustration, the multiplex strategy has been initially described for the situation where there are two nucleotides of interest in a reference sequence and only four probes in an array. Of course, the strategy can be extended to analyze any number of nucleotides in a target sequence by using additional probes. In one variation, each pair of interrogation positions is read from a unique group of four probes. In a block variation, different groups of four probes exhibit the same segment of complementarity with the reference sequence, but the interrogation positions move within a block.

The block and standard multiplex tiling variants can of course be used in combination for different regions of a reference sequence. Either or both variants can also be used in combination with any of the other tiling strategies described.

#### **6.8.3.1.2.7. Helper Mutations**

Occasionally small regions of a reference sequence give a low hybridization signal as a result of annealing of probes.

The self-annealing reduces the amount of probe effectively available for hybridizing to the target. Although such regions of the target are generally small and the reduction of hybridization signal is usually not so substantial as to obscure the sequence of this region, this concern can be avoided by the use of probes incorporating helper mutations.

The helper mutation(s) serve to break-up regions of internal complementarity within a probe and thereby prevent annealing.

Usually, one or two helper mutations are quite sufficient for this purpose. The inclusion of helper mutations can be beneficial in any of the tiling strategies noted above. In general each probe having a particular interrogation position has the same helper mutation(s). Thus, such probes have a segment in common which shows perfect complementarity with a reference sequence, except that the segment contains at least one helper mutation (the same in each of the probes) and at least one interrogation position (different in all of the probes). For example, in the basic tiling strategy, a probe from the first probe set comprises a segment containing an interrogation position and showing perfect

complementarity with a reference sequence except for one or two helper mutations. The corresponding probes from the second, third and fourth probe sets usually comprise the same segment (or sometimes a subsequence thereof including the helper mutation(s) and interrogation position), except that the base occupying the interrogation position varies in each probe.

Usually, the helper mutation tiling strategy is used in conjunction with one of the tiling strategies described above.

The probes containing helper mutations are used to tile regions of a reference sequence otherwise giving low hybridization signal (e.g., because of self-complementarity), and the alternative tiling strategy is used to tile intervening regions.

#### **6.8.3.1.2.8. Pooling Strategies**

Pooling strategies also employ arrays of immobilized probes. Probes are immobilized in cells of an array, and the hybridization signal of each cell can be determined independently of any other cell. A particular cell may be occupied by pooled mixture of probes. Although the identity of each probe in the mixture is known, the individual probes in the pool are not separately addressable. Thus, the hybridization signal from a cell is the aggregate of that of the different probes occupying the cell. In general, a cell is scored as hybridizing to a target sequence if at least one probe occupying the cell comprises a segment exhibiting perfect complementarity to the target sequence.

A simple strategy to show the increased power of pooled strategies over a standard tiling is to create three cells each containing a pooled probe having a single pooled position, the pooled position being the same in each of the pooled probes. At the pooled position, there are two possible nucleotides, allowing the pooled probe to hybridize to two target sequences. In tiling terminology, the pooled position of each probe is an interrogation position. As will become apparent, comparison of the hybridization intensities of the pooled probes from the three cells reveals the identity of the nucleotide in the target sequence corresponding to the interrogation position (i.e., that is matched with the

interrogation position when the target sequence and pooled probes are maximally aligned for complementarity).

The three cells are assigned probe pools that are perfectly complementary to the target except at the pooled position, which is occupied by a different pooled nucleotide in each probe.

With 3 pooled probes, all 4 possible single base pair states (wild and 3 mutants) are detected. A pool hybridizes with a target if some probe contained within that pool is complementary to that target.

A cell containing a pair (or more) of oligonucleotides lights up when a target complementary to any of the oligonucleotide in the cell is present. Using the simple strategy, each of the four possible targets (wild and three mutants) yields a unique hybridization pattern among the three cells.

Since a different pattern of hybridizing pools is obtained for each possible nucleotide in the target sequence corresponding to the pooled interrogation position in the probes, the identity of the nucleotide can be determined from the hybridization pattern of the pools. Whereas, a standard tiling requires four cells to detect and identify the possible single-base substitutions at one location, this simple pooled 45 strategy only requires three cells.

A more efficient pooling strategy for sequence analysis is the 'Trellis' strategy. In this strategy, each pooled probe has a segment of perfect complementarity to a reference sequence except at three pooled positions. One pooled position is an N pool. The three pooled positions may or may not be contiguous in a probe. The other two pooled positions are selected from the group of three pools consisting of (1) M or K, (2) R or Y and (3) W or S, where the single letters are IUPAC standard ambiguity codes. The sequence of a pooled probe is thus, of the form XXXN[(M/K) or (R/Y) or (W/S)][(M/K) or (R/Y) or (W/S)]XXXXXX, where XXX represents bases complementary to the reference sequence. The three pooled positions may be in any order, and may be contiguous or separated by intervening nucleotides. For, the two positions occupied by [(M/K) or (R/Y) or (W/S)], two choices must be made. First, one must select one of the following three pairs of

pooled nucleotides (1) M/K, (2) R/Y and (3) W/S. The one of three pooled nucleotides selected may be the same or different at the two pooled positions. Second, supposing, for example, one selects M/K at one position, one must then choose between M or K. This choice should result in selection of a pooled nucleotide comprising a nucleotide that complements the corresponding nucleotide in a reference sequence, when the probe and reference sequence are maximally aligned. The same principle governs the selection between R and Y, and between W and S. A trellis pool probe has one pooled position with four possibilities, and two pooled positions, each with two possibilities. Thus, a trellis pool probe comprises a mixture of 16 ( $4 \times 2 \times 2$ ) probes. Since each pooled position includes one nucleotide that complements the corresponding nucleotide from the reference sequence, one of these 16 probes has a segment that is the exact complement of the reference sequence. A target sequence that is the same as the reference sequence (i.e., a wildtype target) gives a hybridization signal to each probe cell. Here, as in other tiling methods, the segment of complementarity should be sufficiently long to permit specific hybridization of a pooled probe to a reference sequence be detected relative to a variant of that reference sequence. Typically, the segment of complementarity is about 9-21 nucleotides.

A target sequence is analyzed by comparing hybridization intensities at three pooled probes, each having the structure described above. The segments complementary to the reference sequence present in the three pooled probes show some overlap.

Sometimes the segments are identical (other than at the interrogation positions). However, this need not be the case.

For example, the segments can tile across a reference sequence in increments of one nucleotide (i.e., one pooled probe differs from the next by the acquisition of one nucleotide at the 5' end and loss of a nucleotide at the 3' end). The three interrogation positions may or may not occur at the same relative positions within each pooled probe (i.e., spacing from a probe terminus). All that is required is that one of the three interrogation positions from each of the three pooled probes aligns with the same nucleotide in the reference sequence, and that this interrogation position is occupied by a different pooled nucleotide in each of the three probes. In one of the three probes, the

interrogation position is occupied by an N. In the other two pooled probes the interrogation position is occupied by one of (M/K) or (R/Y) or (W/S).

In the simplest form of the trellis strategy, three pooled probes are used to analyze a single nucleotide in the reference sequence. Much greater economy of probes is achieved when more pooled probes are included in an array.

For example, consider an array of five pooled probes each having the general structure outlined above. Three of these pooled probes have an interrogation position that aligns with the same nucleotide in the reference sequence and are used to read that nucleotide. A different combination of three probes have an interrogation position that aligns with a different nucleotide in the reference sequence. Comparison of these three probe intensities allows analysis of this second nucleotide. Still another combination of three pooled probes from the set of five have an interrogation position that aligns with a third nucleotide in the reference sequence and these probes are used to analyze that nucleotide. Thus, three nucleotides in the reference sequence are fully analyzed from only five pooled probes. By comparison, the basic tiling strategy would require 12 probes for a similar analysis.

The trellis strategy employs an array of probes having at least three cells, each of which is occupied by a pooled probe as described above.

Consider the use of three such pooled probes for analyzing a target sequence, of which one position may contain any single base substitution to the reference sequence (i.e, there are four possible target sequences to be distinguished).

Three cells are occupied by pooled probes having a pooled interrogation position corresponding to the position of possible substitution in the target sequence, one cell with an 'N', one cell with one of 'M' or 'K', and one cell with 'R' or 'Y'. An interrogation position corresponds to a nucleotide in the target sequence if it aligns adjacent with that nucleotide when the probe and target sequence are aligned to maximize complementarity. Note that although each of the pooled probes has two other pooled positions, these positions are not relevant for the present illustration. The positions are only relevant when more than one position in the target sequence is to be read, a

circumstance that will be considered later. For present purposes, the cell with the 'N' in the interrogation position lights up for the wildtype sequence and any of the three single base substitutions of the target sequence.

A further class of strategies involving pooled probes are termed coding strategies. These strategies assign code words from some set of numbers to variants of a reference sequence. Any number of variants can be coded. The variants can include multiple closely spaced substitutions, deletions or insertions. The designation letters or other symbols assigned to each variant may be any arbitrary set of numbers, in any order. For example, a binary code is often used, but codes to other bases are entirely feasible. The numbers are often assigned such that each variant has a designation having at least one digit and at least one nonzero value for that digit.

For example, in a binary system, a variant assigned the number 101, has a designation of three digits, with one possible nonzero value for each digit.

The designation of the variants are coded into an array of pooled probes comprising a pooled probe for each nonzero value of each digit in the numbers assigned to the variants.

For example, if the variants are assigned successive number in a numbering system of base  $m$ , and the highest number assigned to a variant has  $n$  digits, the array would have about  $n \times (m - 1)$  pooled probes. In general,  $\log_m (3N+1)$  probes are required to analyze all variants of  $N$  locations in a reference sequence, each having three possible mutant substitutions.

For example, 10 base pairs of sequence may be analyzed with only 5 pooled probes using a binary coding system.

Each pooled probe has a segment exactly complementary to the reference sequence except that certain positions are pooled.

The segment should be sufficiently long to allow specific hybridization of the pooled probe to the reference sequence relative to a mutated form of the reference sequence. As in other tiling strategies, segments lengths of 9-21 nucleotides are typical. Often the probe



has no nucleotides other than the 9-21 nucleotide segment. The pooled positions comprise nucleotides that allow the pooled probe to hybridize to every variant assigned a particular nonzero value in a particular digit. Usually, the pooled positions further comprises a nucleotide that allows the pooled probe to hybridize to the reference sequence. Thus, a wildtype target (or reference sequence) is immediately recognizable from all the pooled probes being lit.

When a target is hybridized to the pools, only those pools comprising a component probe having a segment that is exactly complementary to the target light up. The identity of the target is then decoded from the pattern of hybridizing pools. Each pool that lights up is correlated with a particular value in a particular digit. Thus, the aggregate hybridization patterns of each lighting pool reveal the value of each digit in the code defining the identity of the target hybridized to the array.

#### **6.8.3.1.2.9. Bridging Strategy**

Probes that contain partial matches to two separate (i.e., non contiguous) subsequences of a target sequence sometimes hybridize strongly to the target sequence. In certain instances, such probes have generated stronger signals than probes of the same length which are perfect matches to the target sequence. It is believed (but not necessary to the invention) that this observation results from interactions of a single target sequence with two or more probes simultaneously. This invention exploits this observation to provide arrays of probes having at least first and second segments, which are respectively complementary to first and second subsequences of a reference sequence. Optionally, the probes may have a third or more complementary segments. These probes can be employed in any of the strategies noted above.

The two segments of such a probe can be complementary to disjoint subsequences of the reference sequences or contiguous subsequences. \* If the latter, the two segments in the probe are inverted relative to the order of the complement of the reference sequence. The two subsequences of the reference sequence each typically comprises about 3 to 30 contiguous nucleotides. The subsequences of the reference sequence are sometimes separated by 0, 1, 2 or 3 bases. Often the sequences, are adjacent and nonoverlapping.

The bridging strategy offers the following advantages:

(1) Higher discrimination between matched and mismatched probes, (2) The possibility of using longer probes in a bridging tiling, thereby increasing the specificity of the hybridization, without sacrificing discrimination, (3) The use of probes in which an interrogation position is located very off-center relative to the regions of target complementarity. This may be of particular advantage when, for example, when a probe centered about one region of the target gives low hybridization signal. The low signal is overcome by using a probe centered about an adjoining region giving a higher hybridization signal. (4) Disruption of secondary structure that might result in annealing of certain probes (see previous discussion of helper mutations).

#### 6.8.3.1.2.10. Deletion Tiling

Deletion tiling is related to both the bridging and helper mutant strategies described above. In the deletion strategy, comparisons are performed between probes sharing a common deletion but differing from each other at an interrogation position located outside the deletion. For example, a first probe comprises first and second segments, each exactly complementary to respective first and second subsequences of a reference sequence, wherein the first and second subsequences of the reference sequence are separated by a short distance (e.g., 1 or 2 nucleotides). The order of the first and second segments in the probe is usually the same as that of the complement to the first and second subsequences in the reference sequence.

Such tilings sometimes offer superior discrimination in hybridization intensities between the probe having an interrogation position complementary to the target and other probes. Thermodynamically, the difference between the hybridizations to matched and mismatched targets for the probe set shown above is the difference between a single-base bulge, and a large asymmetric loop (e.g., two bases of target, one of probe). This often results in a larger difference in stability than the comparison of a perfectly matched probe with a probe showing a single base mismatch in the basic tiling strategy.

The use of deletion or bridging probes is quite general. These probes can be used in any of the tiling strategies of the invention. As well as offering superior discrimination, the use of deletion or bridging strategies is advantageous for certain probes to avoid self-hybridization (either within a probe or between two probes of the same sequence)

#### 6.8.3.1.3. Preparation of Target Samples

The target polynucleotide, whose sequence is to be determined, is usually isolated from a tissue sample. If the target is genomic, the sample may be from any tissue (except exclusively red blood cells). For example, whole blood, peripheral blood lymphocytes or PBMC, skin, hair or semen are convenient sources of clinical samples. These sources are also suitable if the target is RNA. Blood and other body fluids are also a convenient source for isolating viral nucleic acids. If the target is mRNA, the sample is obtained from a tissue in which the mRNA is expressed. If the polynucleotide in the sample is RNA, it is usually reverse transcribed to DNA. DNA samples or cDNA resulting from reverse transcription are usually amplified, e.g., by PCR. Depending on the selection of primers and amplifying enzyme(s), the amplification product can be RNA or DNA.

Paired primers are selected to flank the borders of a target polynucleotide of interest. More than one target can be simultaneously amplified by multiplex PCR in which multiple paired primers are employed. The target can be labelled at one or more nucleotides during or after amplification. For some target polynucleotides (depending on size of sample), e.g., episomal DNA, sufficient DNA is present in the tissue sample to dispense with the amplification step.

When the target strand is prepared in single-stranded form as in preparation of target RNA, the sense of the strand should of course be complementary to that of the probes on the chip. This is achieved by appropriate selection of primers.

The target is preferably fragmented before application to the chip to reduce or eliminate the formation of secondary structures in the target. The average size of targets segments following hybridization is usually larger than the size of probe on the chip.

### 6.8.3.2. Modes Of Practicing The Invention

In an exemplary method, light is shone through a mask to activate functional (for oligonucleotides, typically an -OH) groups protected with a photoremovable protecting group on a surface of a solid support. After light activation, a nucleoside building block, itself protected with a photoremovable protecting group (at the 5'-OH), is coupled to the activated areas of the support. The process can be repeated, using different masks or mask orientations and building blocks, to prepare very dense arrays of many different oligonucleotide probes. New methods for the combinatorial chemical synthesis of peptide, polycarbamate, and oligonucleotide arrays have recently been reported (see Fodor et al., 1991, *Science* 251: 767-773; Cho et al., 1993, *Science* 261: 1303-1305; and Southern et al., 1992, *Genomics* 13: 1008-10017, each of which is incorporated herein by reference). These arrays, or biological chips (see Fodor et al., 1993, *Nature* 364: 555-556, incorporated herein by reference), harbor specific chemical compounds at precise locations in a high-density, information rich format, and are a powerful tool for the study of biological recognition processes. A particularly exciting application of the array technology is in the field of DNA sequence analysis. The hybridization pattern of a DNA target to an array of shorter oligonucleotide probes is used to gain primary structure information of the DNA target. This format has important applications in sequencing by hybridization, DNA diagnostics and in elucidating the thermodynamic parameters affecting nucleic acid recognition.

Conventional DNA sequencing technology is a laborious procedure requiring electrophoretic size separation of labeled DNA fragments. An alternative approach, termed Sequencing By Hybridization (SBH), has been proposed (Lysov et al., 1988, *Dokl. Akad. Nauk SSSR* 303:1508-1511; Bains et al., 1988, *J. Theor. Biol.* 135:303-307; and Drmanac et al., 1989, *Genomics* 4:114-128, incorporated herein by reference). This method uses a set of short oligonucleotide probes of defined sequence to search for complementary sequences on a longer target strand of DNA. The hybridization pattern is used to reconstruct the target DNA sequence. It is envisioned that hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA. In immediate applications of this hybridization methodology, a small number of probes can be used to interrogate local DNA sequence.

Hybridization methodology can be carried out by attaching target DNA to a surface. The target is interrogated with a set of oligonucleotide probes, one at a time (see Strezoska et al., 1991, Proc. Natl. Acad. Sci. USA 88:10089-10093, and Drmanac et al., 1993, Science 260:1649-1652, each of which is incorporated herein by reference). This approach can be implemented with well established methods of immobilization and hybridization detection, but involves a large number of manipulations. For example, to probe a sequence utilizing a full set of octanucleotides, tens of thousands of hybridization reactions must be performed. Alternatively, SBH can be carried out by attaching probes to a surface in an array format where the identity of the probes at each site is known. The target DNA is then added to the array of probes.

The hybridization pattern determined in a single experiment directly reveals the identity of all complementary probes.

As noted above, a preferred method of oligonucleotide probe array synthesis involves the use of light to direct the synthesis of oligonucleotide probes in high-density, miniaturized arrays. Photolabile 5'-protected N-acyl-deoxynucleoside phosphoramidites, surface linker chemistry, and versatile combinatorial synthesis strategies have been developed for this technology. Matrices of spatially-defined oligonucleotide probes have been generated, and the ability to use these arrays to identify complementary sequences has been demonstrated by hybridizing fluorescent labeled oligonucleotides to the DNA chips produced by the methods. The hybridization pattern demonstrates a high degree of base specificity and reveals the sequence of oligonucleotide targets.

The surface of a solid support modified with photolabile protecting groups is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at the 5'-hydroxyl with a photolabile group) is then presented to the surface and coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is

presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of products is obtained.

Light directed chemical synthesis lends itself to highly efficient synthesis strategies which will generate a maximum number of compounds in a minimum number of chemical steps.

To carry out hybridization of DNA targets to the probe arrays, the arrays are mounted in a thermostatically controlled hybridization chamber. Fluorescein labeled DNA targets are injected into the chamber and hybridization is allowed to proceed for 5 min to 24 hr. The surface of the matrix is scanned in an epifluorescence microscope (Zeiss Axioscop 20) equipped with photon counting electronics using 50 -100 uW of 488 nm excitation from an Argon ion laser (Spectra Physics Model 2020). Measurements may be made with the target solution in contact with the probe matrix or after washing. Photon counts are stored and image files are presented after conversion to an eight bit image format.

Arrays of oligonucleotides can be efficiently generated by light-directed synthesis and can be used to determine the identity of DNA target sequences. Because combinatorial strategies are used, the number of compounds increases exponentially while the number of chemical coupling cycles increases only linearly. For example, synthesizing the complete set of  $4^8$  (65,536) octanucleotides will add only four hours to the synthesis for the 16 additional cycles.

Furthermore, combinatorial synthesis strategies can be implemented to generate arrays of any desired composition.

For example, because the entire set of dodecamers ( $4^{12}$ ) can be produced in 48 photolysis and coupling cycles ( $b^n$  compounds requires  $b \times n$  cycles), any subset of the dodecamers (including any subset of shorter oligonucleotides) can be constructed with the correct lithographic mask design in 48 or fewer chemical coupling steps. In addition, the number of compounds in an array is limited only by the density of synthesis sites and the overall array size. Recent experiments have demonstrated hybridization to probes synthesized in 25 um sites. At this resolution, the entire set of 65,536 octanucleotides can be placed in an

array measuring 0.64 cm square, and the set of 1,048,576 dodecanucleotides requires only a 2.56 cm array.

Genome sequencing projects will ultimately be limited by DNA sequencing technologies. Current sequencing methodologies are highly reliant on complex procedures and require substantial manual effort. Sequencing by hybridization has the potential for transforming many of the manual efforts into more efficient and automated formats. Light-directed synthesis is an efficient means for large scale production of miniaturized arrays for SBH. The oligonucleotide arrays are not limited to primary sequencing applications. Because single base changes cause multiple changes in the hybridization pattern, the oligonucleotide arrays provide a powerful means to check the accuracy of previously elucidated DNA sequence, or to scan for changes within a sequence. In the case of octanucleotides, a single base change in the target DNA results in the loss of eight complements, and generates eight new complements. Matching of hybridization patterns may be useful in resolving sequencing ambiguities from standard gel techniques, or for rapidly detecting DNA mutational events. The potentially very high information content of light-directed oligonucleotide arrays will change genetic diagnostic testing. Sequence comparisons of hundreds to thousands of different genes will be assayed simultaneously instead of the current one, or few at a time format. Custom arrays can also be constructed to contain genetic markers for the rapid identification of a wide variety of pathogenic organisms.

Oligonucleotide arrays can also be applied to study the sequence specificity of RNA or protein-DNA interactions.

Experiments can be designed to elucidate specificity rules of non Watson-Crick oligonucleotide structures or to investigate the use of novel synthetic nucleoside analogs for antisense or triple helix applications. Suitably protected RNA monomers may be employed for RNA synthesis. The oligonucleotide arrays should find broad application deducing the thermodynamic and kinetic rules governing formation and stability of oligonucleotide complexes.

Other than the use of photoremovable protecting groups, the nucleoside coupling chemistry is very similar to that used routinely today for oligonucleotide synthesis.



## 7 PLANT GENE EXPRESSION

### 7. General Considerations

The present invention provides a chimeric recombinant DNA molecule comprising: a plurality of DNA sequences, each of which comprises a plant-functional promoter linked to a coding region, which encodes a virus-associated coat protein, wherein said DNA sequences are preferably linked in-tandem so that they are expressed in virus-susceptible plant cells transformed with said recombinant DNA molecule to impart resistance to said viruses; as well as methods for transforming plants with the chimeric constructs and for selecting plants which express at least one of said DNA sequences imparting viral resistance

A method for making a genetically modified plant comprising regenerating a whole plant from a plant cell that has been transfected with DNA sequences comprising a first gene whose expression results in an altered plant phenotype linked to a transiently active promoter, the gene and promoter being separated by a blocking sequence flanked on either side by specific excision sequences, a second gene that encodes a recombinase specific for the specific excision sequences linked to a repressible promoter, and a third gene that encodes the repressor specific for the repressible promoter. Also a method for making a genetically modified hybrid plant by hybridizing a first plant regenerated from a plant cell that has been transfected with DNA sequences comprising a first gene whose expression results in an altered plant phenotype linked to a transiently active promoter, the gene and promoter being separated by a blocking sequence flanked on either side by specific excision sequences to a second plant regenerated from a second plant cell that has been transfected with DNA sequences comprising a second gene that encodes a recombinase specific for the specific excision sequences linked to a promoter that is active during seed germination, and growing a hybrid plant from the hybrid seed. Plant cells, plant tissues, plant seed and whole plants containing the above DNA sequences are also claimed.

The present invention is also directed to a DNA construct formed from a fusion gene which includes a trait DNA molecule and a silencer DNA molecule. The trait DNA molecule has a length that is insufficient to impart a desired trait to plants transformed

with the trait DNA molecule. The silencer DNA molecule is operatively coupled to the trait DNA molecule with the trait and silencer DNA molecules collectively having sufficient length to impart the trait to plants transformed with the DNA construct. Expression systems, host cells, plants, and plant seeds containing the DNA construct are disclosed. The present invention is also directed to imparting multiple traits to a plant.

The present invention is also directed to methods of introgressing one or more desired quantitative traits into a plant comprising screening one or more restriction fragment length polymorphisms (RFLP) for association with desired quantitative traits (QT), selecting one or more RFLP's showing association with the desired QT's, developing a mathematical model based on the magnitude of the association of RFLP(s) to predict the degree of expression of the desired QT's, and using the thus-selected RFLP(s) and the mathematical model in a plant breeding program to predict the degree of introgression and expression of the desired QT's in plant progeny.

A method for producing one of the following proteins in transgenic monocot plant cells is disclosed: (i) mature, glycosylated  $\alpha$ -antitrypsin (AAT) having the same N-terminal amino acid sequence as mature AAT produced in humans and a glycosylation pattern which increases serum half-life substantially over that of mature non-glycosylated AAT; (ii) mature, glycosylated antithrombin III (ATIII) having the same N-terminal amino acid sequence as mature ATIII produced in humans; (iii) mature human serum albumin (HSA) having the same N-terminal amino acid sequence as mature HSA produced in humans and having the folding pattern of native mature HSA as evidenced by its bilirubin-binding characteristics; and (iv) mature, active subtilisin BPN' (BPN') having the same N-terminal amino acid sequence as BPN' produced in *Bacillus*. Monocot plants cells are transformed with a chimeric gene which includes a DNA coding sequence encoding a fusion protein having an (i) N-terminal moiety corresponding to a rice  $\alpha$ -amylase signal sequence peptide and, (iii) immediately adjacent the C-terminal amino acid of said peptide, a protein moiety corresponding to the mature protein to be produced.

A process for commercially propagating plants by tissue culture in such a way as both to conserve desired plant morphology and to transform the plant with respect to one or more desired genes. The method includes the steps of (a) creating an Agrobacterium

vector containing the gene sequence desired to be transferred to the propagated plant, preferably together with a marker gene; (b) taking one or more petiole explants from a mother plant and inoculating them with the Agrobacterium vector; (c) conducting callus formation in the petiole sections in culture, in the dark; and (d) culturing the resulting callus in growth medium containing a benzylamino growth regulator such as benzylaminopurine or, most preferably, benzylaminopurine-riboside. Additional optional growth regulators including auxins and cytokinins (indole butyric acid, benzylamine, benzyladenine, benzylaminopurine, alpha naphthylacetic acid and others known in the art) may also be present. Preferably, the petiole tissue is taken from *Pelargonium x domesticum* and the Agrobacterium vector contains an antisense gene for ACC synthase or ACC oxidase to prevent ACC synthase or ACC oxidase expression and, in turn, the ethylene formation for which these enzymes are precursors.

### 7.1 Experimental Approach

This invention is related to the genetic engineering of plants and to a means and method (use of DNA construct) for conferring a plurality of traits, including resistance to viruses, to a plant using a vector encoding a plurality of genes, such as coat protein genes, protease genes, or replicase genes. The field of the invention is plant genetics, including genetic mapping and restriction fragment length polymorphism technology.

The present invention also relates to:

- (i) the production of mature proteins in plant cells, and in particular, to the production of proteins in mature secreted form.
- (i) the development of techniques for the commercial production of transgenic plants.

## **7.2. PRODUCTION OF VIRUS RESISTANT PLANTS**

The genetic manipulation of plants is centuries old, and modern crop yields and disease- and pest- resistances often owe much to traditional plant genetic engineering. Classical plant breeding methods are time-consuming and subject to chance, however, so the recent advent of recombinant DNA techniques is promising. This promise is encouraging especially with respect to enabling plant geneticists to identify and to clone specific genes for desirable traits, and to introduce such genes into already useful varieties of plants.

### **7.2.1 Infection of crops by plant virus**

Many agriculturally important crops are susceptible to infection by plant viruses, which can seriously damage a crop, reduce its economic value to the grower, and increase its cost to the consumer. Attempts to control or prevent infection of a crop by a plant virus have been made, yet viral pathogens continue to be a significant problem in agriculture.

### **7.2.2 Production of virus resistant plants**

Scientists have recently developed means to produce virus resistant plants using genetic engineering techniques. Such an approach is advantageous in that the genetic material which provides the protection is incorporated into the genome of the plant itself and can be passed on to its progeny. A host plant is resistant if it possesses the ability to suppress or retard the multiplication of a virus, or the development of pathogenic symptoms. "Resistant" is the opposite of "susceptible," and may be divided into: (1) high, (2) moderate, or (3) low resistance, depending upon its effectiveness. Essentially, a resistant plant shows reduced or no symptom expression, and virus multiplication within it is reduced or negligible.

Several different types of host resistance to viruses are recognized. The host may be resistant to: (1) establishment of infection, (2) virus multiplication, or (3) viral movement.

### **7.2.3 EXAMPLE OF PLANT VIRUS: POTYVIRUS**

Potyriviruses are a distinct group of plant viruses which are pathogenic to various crops, and which demonstrate cross-infectivity between plant members of different families. Potyriviruses include watermelon mosaic virus-2 (WMV); papaya ringspot virus strains papaya ringspot and watermelon mosaic I (PRV-p and PRV-w), two closely related

members of the plant potyvirus group which were at one time classified as distinct virus types, but are presently classified as different strains of the same virus; zucchini yellow mosaic virus (ZYMV); potato virus Y; tobacco etch and many others.

These viruses consist of flexous, filamentous particles of dimensions approximately 780 X 12 nanometers. The viral particles contain a single-stranded RNA genome containing about 10,000 nucleotides of positive (+, coding, or sense) polarity. Translation of the RNA genome of potyviruses shows that the RNA encodes a single Large polyprotein of about 330 kD. This polyprotein contains several proteins, one of which is a 49kD protease that is specific for the cleavage of the polyprotein into at least six (6) other peptides. These proteins can be found in the infected plant cell and form the necessary components for viral replication. One of the proteins contained within this polyprotein is a 35kD capsid or coat protein which coats and protects the viral RNA from degradation. Another protein is the nuclear inclusion protein, also referred to as replicase, which is believed to function in the replication of the viral RNA. In the course of a potyviral infection, the replicase protein (60 kDa, also referred to as the nuclear inclusion B protein) and the protease protein (50kDa, also referred to as the nuclear inclusion I or nuclear inclusion A protein) are posttranslationally transported across the nuclear membrane into the nucleus of the plant cell at the later stages of viral infection and accumulate to high levels.

Generally, the coat protein gene is located at the 3'-end of the RNA, just prior to a stretch of terminal adenine nucleotide residues (200 to 300 bases). The location of the 49 Kd protease gene appears to be conserved in these viruses. In the tobacco etch virus, the protease cleavage site has been determined to be the dipeptide Gln-Ser, Gln-Gly or Gln-Ala. Conservation of these dipeptides as the cleavage sites in these viral polyproteins is apparent from the sequences of the above-listed potyviruses.

Expression of the coat protein genes from tobacco mosaic virus, alfalfa mosaic virus, cucumber mosaic virus, and potato virus X, among others, in transgenic plants has resulted in plants which are resistant to infection by the respective virus. Some evidence of heterologous protection has also been reported. For example, Namba et al., Phytopathology, 82, 940 (1992) report that expression of coat protein genes from

watermelon mosaic virus-2 or zucchini yellow mosaic virus in transgenic tobacco plants conferred protection against six other potyviruses: bean yellow mosaic virus, potato virus Y, pea mosaic virus, clover yellow vein virus, pepper mottle virus and tobacco etch virus. Stark et al., *Biotechnology*, 1, 1257 (1989) report that expression of the potyvirus soybean mosaic virus in transgenic plants provided protection against two serologically unrelated potyviruses: tobacco etch virus and potato virus Y.

However, expression of a preselected coat protein gene does not reliably confer heterologous protection to a plant. For example, transgenic squash plants containing the CMV-C coat protein gene and which have been shown to be resistant to CMV-C strain, are not protected against several highly virulent strains of CMV, including CW-V-27 and CARNA- 5. Thus, a need exists for improved methods to impart potyvirus resistance to plants.

#### **7.2.4 USING "PATHOGEN DRIVEN RESISTANCE" (PDR) FOR DEVELOPING VIRUS RESISTANT TRANSGENIC PLANTS**

Control of plant virus diseases took a major step forward in the last decade when it was shown in 1986 that the tobacco mosaic virus ("TMV") coat protein gene that was expressed in transgenic tobacco conferred resistance to TMV (Powell-Abel, P., et al., "Delay of Disease Development in Transgenic Plants that Express the Tobacco Mosaic Virus Coat Protein Gene," *Science*, 232:738-43 (1986)). The concept of pathogen-derived resistance ("PDR"), which states that pathogen genes that are expressed in transgenic plants will confer resistance to infection by the homologous or related pathogens (Sanford, J.C., et al. "The Concept of Parasite-Derived Resistance - Deriving Resistance Genes from the Parasite's Own Genome," *J. Theor. Biol.*, 113:395-405 (1985)) was introduced at about the same time. Since then, numerous reports have confirmed that PDR is a useful strategy for developing transgenic plants that are resistant to many different viruses (Lomonosoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," *Ann. Rev. Phytopathol.*, 33:323-43 (1995)).

Only eight years after the report by Beachy and colleagues (Powell-Abel, P., et al., "Delay of Disease Development in Transgenic Plants that Express the Tobacco Mosaic Virus Coat Protein Gene," *Science*, 232:738-43 (1986)), Grunet, R., "Development of

Virus Resistant Plants via Genetic Engineering," *Plant Breeding Reviews*, 12:47-49 (1994) reviewed the PDR literature and listed the successful development of virus resistant transgenic plants to at least 11 different groups of plant viruses.

#### **7.2.1.4.1 Utilizing The Coat Protein Genes**

The vast majority of reports have utilized the coat protein genes of the viruses that are targeted for control. Although the testing of transgenic plants have been largely confined to laboratory and greenhouse experiments, a growing number of reports showed that resistance is effective under field conditions (e.g., Grumet, R., "Development of Virus Resistant Plants via Genetic Engineering," *Plant Breeding Reviews*, 12:47-49 (1994)). Two virus resistant crops have been deregulated by APHIS/USDA and thus are approved for unrestricted release into the environment in the U.S.A. Squash that are resistant to watermelon mosaic virus 2 and zucchini yellow mosaic potyviruses have been commercialized (Fuchs, M., et al., "Resistance of Transgenic Hybrid Squash ZW-20 Expressing the Coat Protein Genes of Zucchini Yellow Mosaic Virus and Watermelon Mosaic Virus 2 to Mixed Infections by Both Potyviruses," *Bio/Technology*, 13:1466-73 (1995); Tricoli, D.M., et al., "Field Evaluation of Transgenic Squash Containing Single or Multiple Virus Coat Protein Gene Constructs for Resistance to Cucumber Mosaic Virus, Watermelon Mosaic Virus 2, and Zucchini Yellow Mosaic Virus," *Bio/Technology*, 13:1458-65 (1995)). Also, a transgenic papaya that is resistant to papaya ringspot virus has been developed (Fitch, M. M. M., et al., "Virus Resistant Papaya Derived from Tissues Bombarded with the Coat Protein Gene of Papaya Ringspot Virus," *Bio/Technology*, 10:1466-72 (1992); Tennant, P.F., et al., "Differential Protection Against Papaya Ringspot Virus Isolates in Coat Protein Gene Transgenic Papaya and Classically Cross- Protected Papaya," *Phytopathology*, 84:1359-66 (1994)). This resistant transgenic papaya was recently deregulated by USDA/APHIS. Deregulation of the transgenic papaya is timely, because Hawaii's papaya industry is being devastated by papaya ringspot virus. Undoubtedly, more crops will be deregulated and commercialized in the near future.

#### **7.2.1.4.2 Other effective viral genes**

Interestingly, remarkable progress has been made in developing virus resistant transgenic plants despite a poor understanding of the mechanisms involved in the various forms of pathogen-derived resistance (Lomonossoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," *Ann.-Rev. Phytopathol.*, 33:323-43 (1995)). Although most reports deal



with the use of coat protein genes to confer resistance, a growing number of reports have shown that viral replicase (Golemboski, D.B., et al., "Plants Transformed with a Tobacco Mosaic Virus Nonstructural Gene Sequence are Resistant to the Virus," Proc. Natl. Acad. Sci. USA, 87:6311-15 (1990)), movement protein (e.g., Beck, D. L., et al., "Disruption of Virus Movement Confers Broad-Spectrum Resistance Against Systemic Infection by Plant Viruses with a Triple Gene Block," Proc. Natl. Acad. Sci. USA, 91:10310-14 (1994)), NIa proteases of potyviruses (e.g., Maiti, I.B., et al., "Plants that Express a Potyvirus Proteinase Gene are Resistant to Virus Infection," Proc. Natl. Acad. Sci. USA, 90:6110-14 (1993)), and other viral genes are effective. This led to the conclusion that any part of a plant viral genome gives rise to PDR. Furthermore, the viral genes can be effective in the translatable and nontranslatable sense forms, and less frequently antisense forms (e.g., Baulcombe, D.C., "Mechanisms of Pathogen-Derived Resistance to Viruses in Transgenic Plants," Plant Cell, 8:1833-44 (1996); Dougherty, W. G., et al., "Transgenes and Gene Suppression: Telling us Something New?," Current Opinion in Cell Biology, 7:399- 05 (1995); Lomonossoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," Ann. Rev. Photopathol. 33:323-43 (1995)).

#### 7.2.1.5 RNA-MEDIATED RESISTANCE

##### 7.1.2.1.5.1 Description (A Form of PDR)

RNA-mediated resistance is the form of PDR where there is clear evidence that viral proteins do not play a role in conferring resistance to the transgenic plant. The first clear cases for RNA-mediated resistance were reported in 1992 for tobacco etch ("TEV") potyvirus (Lindbo, et al., "Pathogen-Derived Resistance to a Potyvirus Immune and Resistance Phenotypes in Transgenic Tobacco Expressing Altered Forms of a Potyvirus Coat Protein Nucleotide Sequence," Mol. Plant Microbe Interact., 5:144-53 (1992)), for potato virus Y ("PVY") potyvirus by Van Der Vlugt, R.A.A., et al., "Evidence for Sense RNA-Mediated Protection to PVY in Tobacco Plants Transformed with the Viral Oat Protein Cistron," Plant Mol. Biol., 20:631-39 (1992), and for tomato spotted wilt ("TSWV") tospovirus by de Haan, P., et al., "Characterization of RNA-Mediated Resistance to Tomato Spotted Wilt Virus in Transgenic Tobacco Plants," Bio/Technology, 10:1133-37 (1992). others confirmed the occurrence of RNA-mediated resistance with potyviruses (Smith, H.A., et al., "Transgenic Plant Virus Resistance Mediated by Untranslatable Sense RNAs: Expression, Regulation, and Fate of Nonessential RNAs,"

Plant Cell, 6:1441-53 (1994)), potexviruses (Mueller, E., et al., "Homology-Dependent Resistance: Transgenic Virus Resistance in Plants Related to Homology-Dependent Gene Silencing," Plant Journal, 7:1001-13 (1995)), and TSWV and other tospoviruses (Pang, S.Z., et al., "Resistance of Transgenic *Nicotiana Benthamiana* Plants to Tomato Spotted Wilt and Impatiens Necrotic Spot Tospoviruses: Evidence of Involvement of the N Protein and N Gene RNA in Resistance," Phytopathology, 84:243-49 (1994); Pang, S.-Z., et al., "Different Mechanisms Protect Transgenic Tobacco Against Tomato Spotted Wilt Virus and Impatiens Necrotic Spot Tospoviruses," Bio/Technology 11:819-24 (1993)). More recent work has shown that RNA-mediated resistance also occurs with the comovirus cowpea mosaic virus (Sijen, T., et al., "RNA-Mediated Virus Resistance: Role of Repeated Transgene and Delineation of Targeted Regions," Plant Cell, 8:2227-94 (1996)) and squash mosaic virus (Jan, F.-J., et al., "Genetic and Molecular Analysis of Squash Plants Transformed with Coat Protein Genes of Squash Mosaic Virus," Phytopathology, 86:S16-17 (1996)).

#### 7.2.1.5.2 The Mechanism(S) of RNA-Mediated Resistance

Major advances towards understanding the mechanism(s) of RNA-mediated resistance were made by Dougherty and colleagues in a series of experiments with TEV and PVY. Using TEV, Lindbo, J.A., "Pathogen-Derived Resistance to a Potyvirus Immune and Resistant Phenotypes in Transgenic Tobacco Expressing Altered Forms of a Potyvirus Coat Protein Nucleotide Sequence," Mol. Plant Microbe Interact., 5:144-53 (1992) and Lindbo, J.A., et al., "Untranslatable Transcripts of the Tobacco Etch Virus Coat Protein Gene Sequence can Interfere with Tobacco Etch Virus Replication in Transgenic Plants and Protoplasts," Virology, 189:725-33 (1992) showed that transgenic plants expressing translatable full length coat protein, truncated translatable coat protein, antisense coat protein genes, and nontranslatable coat protein gene had various phenotypic reactions after inoculation with TEV. Transgenic plants displayed resistance, recovery (inoculated plants initially show systemic infection but younger leaves that develop later are symptomless and resistant to the virus), or susceptible phenotypes. Furthermore, they showed that leaves of resistant plants and asymptomatic leaves of recovered plants had relatively low levels of steady state RNA when compared to those in leaves of susceptible plants (Lindbo, J.A., et al., "Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance," Plant Cell, 5:1749-

59 (1993)). However, nuclear run off experiments showed that those plants with low levels of steady state RNA had higher transcription rates of the viral transgene than those plants that were susceptible (and had high steady state RNA levels). To account for these observations, it was proposed "that the resistant state and reduced steady state levels of transgene transcript accumulation are mediated at the cellular level by a cytoplasmic activity that targets specific RNA sequences for inactivation," (Lindbo, J.A., et al., "Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance," Plant Cell, 5:1749-59 (1993)). It was also suggested that the low steady state RNA levels may be due to post-transcriptional gene silencing, a phenomenon that was first proposed by de Carvalho, F., et al., "Suppression of beta-1,3-glucanase Transgene Expression in Homozygous Plants," EMBO J., 11:2595-602 (1992) for the suppression of P-1,3-glucanase transgene in homozygous transgenic plants.

An RNA threshold model was proposed to account for the observations (Lindbo, J.A., et al., "Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance," Plant Cell, 5:1749-59 (1993)). Basically, the model states that there is a cytoplasmic cellular degradation mechanism that acts to limit the RNA levels in plant cells, and that this mechanism is activated when the transgenic RNA transcript goes above a threshold level. The degradation mechanism is specific for the transcript that goes above the threshold level; and if the transcripts that go above a certain threshold is a viral transgene, the virus resistance state is observed in the plant, because the degradation mechanism also targets, for inactivation, the specific sequences of the incoming virus. The model also accounts for the 'recovery' of transgenic plants by suggesting that viral RNA from the systemically invading virus triggers the phenomenon in some transgenic plants that have two copies of the transgenes. Plants that had more than three copies of the transgenes caused the threshold level to be surpassed without the invasion of virus (Goodwin, J., et al., "Genetic and Biochemical Dissection of Transgenic RNA-Mediated Virus Resistance," Plant Cell, 8:95-105 (1996); Smith, H.A., et al., "Transgenic Plant Virus Resistance Mediated by Untranslatable Sense RNAs: Expression, Regulation, and Fate of Nonessential RNAs," Plant Cell, 6:1441-53 (1994)). Although the degradation mechanism is not clear, it is proposed that a cellular RNA dependent RNA polymerase ("RdRp") binds to the transcript

and produces small fragments of antisense RNA which then bind to other transcripts to form duplexes which are then degraded by nucleases that specifically recognize RNA-RNA duplexes. This degradation mechanism is sequence specific, which accounts for the specificity of RNA-mediated resistance.

Work on PVX by Baulcombe and colleagues (English, J. J., et al., "Suppression of Virus Accumulation in Transgenic Plants Exhibiting Silencing of Nuclear Genes," Plant Cell, 8: 179-88 (1996); Mueller, E., et al., "Homology-Dependent Resistance: Transgenic Virus Resistance in Plants Related to Homology-Dependent Gene Silencing," Plant Journal, 7:1001-13 (1995)) confirmed and extended the results by Dougherty and colleagues. An aberrant RNA model which is a modification of the RNA threshold model proposed by Dougherty and colleagues was proposed. The features of the model are similar to Dougherty's except that it states that the RNA level is not the sole trigger to activate the cellular degradation mechanism, but instead aberrant RNA that are produced during the transcription of the transgene plays an important part in activating the cytoplasmic cellular mechanism that degrades specific RNA. The production of aberrant RNA may be enhanced by positional effects of the transgene on the chromosome and by methylation of the transgene DNA. The precise nature of the aberrant RNA is not defined, but it may contain a characteristic that makes it a preferred template for the production of antisense RNA by the host encoded RdRp (Baulcombe, D.C., "Mechanisms of Pathogen-Derived Resistance to Viruses in Transgenic Plants," Plant Cell, 8:1833-44 (1996); English, J. J., et al., "Suppression of Virus Accumulation in Transgenic Plants Exhibiting Silencing of Nuclear Genes," Plant Cell, 8: 179-88 (1996)). Thus, the model also proposes that RdRp and antisense molecules are involved in the degradation mechanism. Baulcombe and colleagues confirmed that plants which show low steady state transgene levels have multiple copies of transgenes and that the low steady state RNA and the accompanying resistant state is due to post-transcriptional gene silencing. The term homology-dependent resistance was proposed to describe the resistance in plants that show homology-dependent gene silencing (Mueller, E., et al., "Homology-Dependent Resistance: Transgenic Virus Resistance in Plants Related to Homology-Dependent Gene Silencing," Plant Journal, 7:1001-13 (1995)).

Experiments with TSWV tospovirus (Pang, S.Z., et al., "Post-Transcriptional Transgene Silencing and Consequent Tospovirus Resistance in Transgenic Lettuce are Affected by Transgene Dosage and Plant Development," Plant Journal, 9:899-09 (1996); Prins, M., et al., "Engineered RNA Mediated Resistance to Tomato Spotted Wilt Virus is Sequence Specific," Mol. Plant Microbe Interact., 9:416-18 (1996)) and cowpea mosaic comovirus (Sijen, T., et al., "RNA-Mediated Virus Resistance: Role of Repeated Transgene and Delineation of Targeted Regions," Plant Cell, 8:2227-94 (1996)) also showed that resistance in transgenic plants is a consequence of post-transcriptional gene silencing. Pang, S.Z., et al., "Post-Transcriptional Transgene Silencing and Consequent Tospovirus Resistance in Transgenic Lettuce are Affected by Transgene Dosage and Plant Development," Plant Journal, 9:899-09 (1996) showed that post-transcriptional gene silencing in transgenic lettuce expressing the N 1S gene of TSWV was influenced by gene dosage and by the developmental stage of the plant. The effect of developmental stage on post-transcriptional gene silencing of transgenes and their effect on resistance had not been previously shown for transgenic plants expressing viral genes, but had been shown to occur in plants expressing other transgenes (de Carvalho, F., et al., "Suppression of beta-1,3-glucanase Transgene Expression in Homozygous Plants," EMBO J., 11:2595-02 (1992)). Post-transcriptional gene silencing could also account for the correlation of low steady state level of N gene RNA in transgenic tobacco showing very high but specific resistance (Pang, S.Z., et al., "Different Mechanisms Protect Transgenic Tobacco Against Tomato Spotted Wilt and Impatiens Necrotic Spot Tospoviruses," Bio/Technology, 11:819-24 (1993)). Prins, M., et al., "Engineered RNA-Mediated Resistance to Tomato Spotted Wilt Virus is Sequence Specific," Molecular Plant Microbe Interactions, 9:416-18 (1996) also reported that post-transcriptional gene silencing occurred with transgenic tobacco expressing the N gene and nonstructural gene of the mRNA. Interestingly, it was found that tobacco with other parts of the TSWV genome were not resistant. They suggested, as one explanation, that those gene fragments which did not confer resistance may not fit the criteria for inducing post-transcriptional gene silencing. Sijen, T., et al., "RNA-Mediated Virus Resistance: Role of Repeated Transgene and Delineation of Targeted Regions," Plant Cell, 8:2227-94 (1996) showed that resistance of transgenic plants expressing the movement protein, replicase, or coat protein were due to post-transcriptional gene silencing. This data also suggested that the 31 region of the movement protein transgene mRNA is the initial target of the silencing mechanism.

The present invention is directed to producing improved disease resistant plants.

## **7.2.2 CREATING TRANSGENIC PLANTS WITH CONTROLLABLE GENES**

This invention also relates to certain transgenic plants and involves a method of creating transgenic plants with controllable genes. More particularly, the invention relates to transgenic plants that have been modified such that expression of a desired introduced gene can be limited to a particular stage of plant development, a particular plant tissue, particular environmental conditions, or a particular time or location, or a combination of these situations.

### **7.2.2.1 Inducible gene promoter: "gene switch"**

Various gene expression control elements that are operable in one or more species of organisms are known. For example, PCT Application WO 90/08826 (Bridges, et al.) discloses an inducible gene promoter that is responsive to an exogenous chemical inducer, called a "gene switch." This promoter can be linked to a gene and introduced into a plant. The gene can be selectively expressed by application of the chemical inducer to activate the promoter directly.

PCT application WO 94/03619 (Bright, et al.) discloses a gene cascade consisting of a gene switch linked to a repressor gene and a repressible operator linked to a disrupter protein capable of disrupting plant development. Growth of the plant can be controlled by the application or withholding of a chemical inducer. While the inducer is present, the repressor is expressed, the promoter attached to the disrupter gene is repressed, the disrupter protein is not expressed, thereby allowing the plant to grow normally. If the chemical inducer is withheld, the gene switch is turned off, the repressible promoter is not repressed, so the disrupter protein is expressed and plant development is disrupted. This system is said to be useful for controlling the escape of plants into the wild by making their continued growth and development dependent on the continued application of a chemical inducer, and to mitigate the problem of preharvest sprouting of grains by withholding the chemical inducer at the last stages of seed development.

#### **7.2.2.2 Tetracycline-controlled plant-active repressor-operator system**

Gatz and Quail (1988) and Gatz, et al. (1992), (Hoppe-Seyler), 372:659-660 (1991), disclose a plant-active repressor-operator system that is controlled by the application of tetracycline. The system consists of the Tn10 tet repressor gene, and a

cauliflower mosaic virus (CaMV) 35S promoter, modified to contain two tet operons and linked to the chloramphenicol acetyltransferase (cat) gene (Gatz and Quail, 1988), or modified to contain three tet operons and linked to the beta-glucuronidase (gus) gene (Gatz, et al., 1992). So long as the Tn10 tet repressor gene is active, the modified promoter is repressed by the interaction of the repressor with the tet operons, and the cat or gus gene is not expressed. The presence of tetracycline inhibits repressor binding, enabling expression of the cat or gus gene.



### **7.2.3 RECOMBINANT PRODUCTION OF PROTEINS**

A major commercial focus of biotechnology is the recombinant production of proteins, including both industrial enzymes and proteins that have important therapeutic uses.

#### **7.2.3.1 Recombinant proteins produced in microbial hosts**

Therapeutic proteins are commonly produced recombinantly by microbial expression systems, such as in *E. coli* and the yeast system *S. cerevisiae*. To date, the cost of recombinant proteins produced in a microbial host has limited the availability of a variety of therapeutically important proteins, such as human serum albumin (HSA) and  $\alpha_1$ -antitrypsin (AAT), to the extent that the proteins are in short supply.

##### **7.2.31.1 Inability of microbial systems to glycosylate (properly) mammalian proteins**

Some therapeutic proteins appear to rely on glycosylation for optimal activity or stability, and the general inability of microbial systems to glycosylate or properly glycosylate mammalian proteins has also limited the usefulness of these recombinant expression systems. In some cases, proper protein folding cannot take place, because of the need for mammalian-specific foldases or other folding conditions.

##### **7.2.3.1.2 Cost per weight ratio and other problems of mammalian expression systems**

To some extent, protein expression in cultured mammalian cells, or in transgenic animals may overcome the limitations of microbial expression systems. However, the cost per weight ratio of the protein is still high in mammalian expression systems, and the risk of protein contamination by mammalian viruses may be a significant regulatory problem. Protein production by transgenic animals also carries the risk of genetic variation from one generation to another. The attendant risk is variation in the recombinant protein produced, for example, variation in protein processing to yield a native active protein with different N-terminal residue.

#### **7.2.3.2 Alternative protein expression system to overcome problems of microbial and mammalian systems**

It would therefore be desirable to produce selected therapeutic and industrial proteins in a protein expression system that largely overcomes problems associated with

microbial and mammalian-cell systems. In particular, production of the proteins should allow large volume production at low cost, and yield properly processed and glycosylated proteins. The production system should also have a relatively stable genotype from generation to generation. These aims are achieved, in the present invention, for the therapeutic proteins AAT, HSA, and antithrombin III (ATIII), and the industrial enzyme subtilisin BPN'.

#### **7.2.3.2.1 Use: Human $\alpha_1$ -antitrypsin**

Human  $\alpha_1$ -antitrypsin (AAT) is a monomer with a molecular weight of about 52kd.

Normal AAT contains 394 residues, with three complex oligosaccharide units exposed to the surface of the molecule, linked to asparagines 46, 83, and 247 (Carrell, P., et al., Nature (1992) 298:329).

AAT is the major plasma proteinase inhibitor whose primary function is to control the proteolytic activity of trypsin, elastase, and chymotrypsin in plasma. In particular, the protein is a potent inhibitor of neutrophil elastase, and a deficiency of AAT has been observed in a number of patients with chronic emphysema of the lungs. A proportion of individuals with serum deficiency of AAT may progress to cirrhosis and liver failure (e.g., Wu, Y., et al., BioEssays 13(4):163 (1991).

Because of the key role of AAT as an elastase inhibitor, and because of the prevalence of genetic diseases resulting in deficient serum levels of AAT, there has been an active interest in recombinant synthesis of AAT, for human therapeutic use. To date, this approach has not been satisfactory for AAT produced by recombinant methods, for the reasons discussed above.

#### **7.2.3.2.2 Use: Human Antithrombin III**

Antithrombin III (ATIII) is the major inhibitor of thrombin and factor Xa, and to a lesser extent, other serine proteases, generated during the coagulation process, e.g., factors IXa, XIa, and XIIa. The inhibitory effect of ATIII is accelerated dramatically by heparin. In patients with a history of deep vein thrombosis and pulmonary embolism, the prevalence of ATIII deficiency is 2- 3%.

ATIII protein has been useful in treating hereditary ATIII deficiency and has wide clinical applications for the prevention of thrombosis in high risk situations, such as surgery and delivery, and for treating acute thrombotic episodes, when used in combination with heparin.

ATIII is a glycoprotein with a molecular weight of 58,200, having 432 amino acids and containing three disulfide linkages and four asparagine-linked biantennary carbohydrate chains. Because of the key role of ATIII as an anti-thrombotic agent, and because of the broad clinical potential in anti-thrombosis therapy, there has been an active interest in recombinant synthesis of ATIII, for human therapeutic use. To date, this approach has not been satisfactory for ATIII produced by microbial or mammalian recombinant methods, for the reasons discussed above.

#### **7.2.3.2.3 Use: Human Serum Albumin**

Serum albumin is the main protein component of plasma. Its main function is regulation of colloidal osmotic pressure in the bloodstream. Serum albumin binds numerous ions and small molecules, including  $\text{Ca}^{2+}$ ,  $\text{Na}^{+}$ ,  $\text{K}^{+}$ , fatty acids, hormones, bilirubin and certain drugs.

Human serum albumin (HSA) is expressed as a 609 amino acid prepro- protein which is further processed by removal of an amino-terminal peptide and an additional six amino acid residues to form the mature protein. The mature protein found in human serum is a monomeric, unglycosylated protein 585 amino acids in length (66 kDal), with a globular structure maintained by 17 disulfide bonds. The pattern of disulfide links forms a structural unit of one small and two large disulfide-linked double loops (Geisow, M.J. et al. (1977) *Biochem. J.* 163:477-484) which forms a high-affinity bilirubin binding site.

HSA is used to expand blood volume and raise low blood protein levels in cases of shock, trauma, and post-surgical recovery. HSA is often administered in emergency situations to stabilize blood pressure.

Because of the key role of HSA as an osmotic stabilizing agent, and because of its broad clinical potential in, e.g., plasma replacement therapy, there has been an active interest in recombinant synthesis of HSA for human therapeutic use. This approach has not been satisfactory for HSA produced by microbial or mammalian recombinant methods, for the reasons discussed above.

#### **7.2.3.2.4 Use: Subtilisin BPN'**

Subtilisin BPN' (BPN') is an important industrial enzyme, particularly for use as a detergent enzyme. Several groups have reported amino acid substitution modifications of the enzyme that are effective in enhancing the activity, pH optimum, stability and/or therapeutic use of the enzyme.

BPN' is expressed in as a 381 amino acid preproenzyme, including 35 amino acid sequence required for secretion and a 77 amino acid moiety which serves as a chaperon to facilitate folding. Studies indicate that the pro moiety acts in trans outside of cells.

To date, large-scale production of BPN' is predominantly by microbial fermentation, which has relatively high costs associated with it. In addition, the enzyme tends to auto-degrade at optimal fermentation growth-medium conditions.

## 7.2.4 PRACTICAL METHOD FOR THE COMMERCIAL PRODUCTION OF TRANSGENIC PLANTS

Translating genetic engineering theory into practice, however, and then furthermore into a commercially practical reality, requires ingenuity. Gene transplantation in plants has already been accomplished at this writing--and examples are cited below--but heretofore no practical method for the commercial production of transgenic plants has been perfected.

### 7.2.4.1 Tissue culture propagation

Apart from the transgenic plant technology per se, it is known to propagate plants by replicating plant cells in culture, or "tissue culture." An early motivating force in the development of tissue culture was the desire to improve upon the relatively slow and low yields of vegetative propagation with the quick and exponential proliferation of new plants from cell culture. Tissue culture methods are made possible by the plant physiological phenomenon of callus formation. When a plant is wounded, a patch of soft cells called a calli grows over the wound and, with time, phenolic compounds accumulate in the soft cells and harden, effectively sealing the wound. While hardened callus is the plant equivalent of scar tissue, callus is different from mammalian scar tissue with respect to its regenerative properties. If a piece of young, still-soft callus is removed and placed in a culture medium containing salts, sugars, vitamins, amino acids and the appropriate plant growth hormones, rather than harden, the cells will continue to divide and give rise to a disorganized mass of undifferentiated cells called a "callus culture." Plant or seedling "explants," or tissue samples, will likewise grow into similar cell cultures. The cultured cells can further be induced to redifferentiate into shoots, roots or whole plants by further culturing with the necessary hormones and growth media.

One of the most serious drawbacks with tissue culture propagation techniques has been the morphologic variation from generation to generation, a problem which is particularly notable in certain species and varieties. For example, as reported in Cassells, A.C., and Carney, B.F., "Adventitious regeneration in Pelargonium x domesticum Bailey," Acta Horticulturae, 212(11), 419-425 (1987), in stem and petiole tissue cultures of Grand Slam (as an example of P. domesticum, also known as Regal Pelargoniums or "Martha Washington" geraniums), up to 16% of the adventitious regenerants were

variants, depending on the explant origin. The authors concluded that genome instability in Grand Slam and presumably other P. domesticum varieties may produce useful variation but mitigates against the use of adventitious regeneration in micropropagation.

The findings of Cassells et al. are consistent with the earlier work of Skirvin, R.M. and Janick, Jules, "Tissue Culture-Induced variation in Scented Pelargonium ssp.," J. Amer. Soc. Hort. Sci. , 101 (3), 281-290 (1976). Skirvin et al. compared tissue culture propagated Pelarcronium plants (from root cuttings, petiole cuttings or calliclones) with plants derived from vegetative propagation, i.e., stem cuttings. The plants derived from stem cuttings were all uniform and identical to the parental clone, whereas those from the root cuttings, petiole cuttings or calliclones were all morphologically distinct with the degree of variability depending upon the cultivar. The authors conclude that the variability associated with calliclones derived from tissue culture is a pool on which selection can be imposed, implying conversely that tissue culturing of this type is inappropriate for use in attempting reliable regeneration of Pelarcronium x domesticum varieties.

Other varieties and species, besides Pelargonium x domesticum, are known and/or believed to suffer morphologic variation when propagated using tissue culture. It can be easily appreciated that any substantial morphologic variation in propagation is unacceptable for commercial propagation of a desired variety or species. Thus, tissue culture methods are not always acceptable for commercial use, even with the potentially much larger yields achievable as compared with prior art vegetative propagation techniques.

#### 7.2.4.2 Gene transplantation

Apart from tissue culture considerations, gene transplantation in plants has achieved some success at this writing. Gene introduction is generally accomplished with a vector such as Agrobacterium. As this technology developed, it was noted that Crown Gall tumors of plants arose at the site of infection of some species of the bacterium Agrobacterium. The cells of Crown Galls acquire the properties of independent, unregulated growth. In culture, such transformed cells grow in the absence of the plant hormones usually necessary for plant cell growth, and the cells retain the transformed phenotype even in the absence of the bacterium. The tumor-inducing agent in

Agrobacterium is a plasmid that integrates some of its DNA into the chromosome of the host plant cells. Ti (tumor-inducing) plasmids exist in Agrobacterium cells as independently replicating genetic units.

Ti plasmids are maintained in Agrobacterium because part of the plasmid DNA, the T-DNA, carries the genes coding for the synthesis of amino acids called opines. The infected plant cell is induced to synthesize these amino acids, but the plant cannot use these amino acids. The Ti plasmid is believed to carry genes coding for enzymes that can degrade opines. Thus, Ti plasmids both make and degrade opines, within the plant cell, which the plant cell cannot metabolically use--presumably giving a selective advantage to the Agrobacterium at least with respect to utilization of the opine metabolites. A second set of genes in T-DNA codes for enzymes which lead to production of hormones which, in turn, cause the infected plant cell to divide in an unregulated way.

In summary terms, T-DNA enters a plant cell by what amounts to the equivalent of bacterial conjugation between the Agrobacterium and the plant cell. In other words, an Agrobacterium organism and a plant cell transfer their DNA in a process analogous to mating. Ultimately, T-DNA becomes incorporated into the genomic plant cell DNA in the plant cell nucleus.

All of the above background illustrates how Agrobacterium species can serve well as vectors for genetic transformation of plant cells. Early gene transfer using Ti plasmids, T-DNA and Agrobacterium was accomplished by the cointegration method, in which T-DNA was first cloned into a standard E. coli cloning vector, and the plant gene was subsequently cloned into a second cloning site carried by the vector. This intermediate vector was introduced into Agrobacterium organisms containing intact Ti plasmids.

Recombination occurred between the homologous regions of the intermediate vector and the wild-type Ti plasmid, and on infection of a plant with the Agrobacterium the recombinant plasmid is transferred to the plant cells.

#### 7.2.4.3 The binary system

Despite the early use of the cointegration method described above--and certainly it still works--the standard method for T-DNA transfer as of this writing is called the "binary system." The binary system was devised when investigators realized that the essential functions for transfer are supplied separately by the T-DNA itself and by the Ti plasmid, and that the components can be carried on separate vectors. The binary vector contains the borders of the T-DNA--needed for excision and integration--and the hormone-producing region of the original T-DNA can be removed and replaced with the foreign gene sequence intended for transfer to the plant cell. One side benefit of the use of binary vectors is that, by removing the hormone-producing regions of the T-DNA, uncontrolled growth of the recipient cells is prevented--or in other words the tumor-causing aspect of the T-DNA is nullified. The vir genes of the Ti plasmid can be supplied on a separate plasmid and etc.; the binary vector technique for gene transfer into plants is well established at this writing.

An example of the use of binary vectors to introduce functional genes into plants came about through experiments to use antisense RNA to control plant gene expression. Early work used binary vectors to introduce antisense polygalacturonase genes into tomato plants, to turn off the polygalacturonase expression which in turn digests pectin, in attempts to reduce bruising of tomato fruit during shipment. The results of these trials were disappointing. However, when binary vectors have been used to transfer antisense ethylene precursor genes into tomato plants, the results have been favorable. The antisense gene prevents expression of the ethylene precursor, no ethylene production occurs during storage of the harvested tomatoes, and thus no ripening occurs until the time ripening is desired, when the fruit can be contacted with ethylene from another source.

Exemplary publications and patents which disclose transgenic plants and various techniques therefor are summarized below.

Pellegrineschi, A., et al., "Improvement of ornamental Characters and Fragrance Production in Lemon-scented Geranium Through Genetic Transformation by Agrobacterium rhizocrenes," Bio/Technology, Vol. 12 (January, 1994) discloses transformation of root cultures by inoculating stem and leaf fragments with Agrobacterium rhizoaenes. An important plasmid in this species of Agrobacterium is the root-inducing



plasmid which can be used to transfer to the plant genome the genes necessary for improved root growth in culture. The use of sterilized petioles as the source of explant material for plant transformation and culture is disclosed.

U.S. Patent No. 5,276,268 to Strauch et al., entitled "Phosphinothricin-Resistance Gene, and Its Use," is directed to the transfer of phosphinothricin-resistance gene into plants using Agrobacterium species. A modification of the binary vector method is discussed, and the phosphinothricin-resistance gene nucleic acid sequences are provided.

U.S. Patent No. 5,283,184 to Jorgenson et al. is entitled "Genetic Engineering of Novel Plant Phenotypes" and discusses transgene formation and propagation in tissue culture, as well mentioning Pelargonium and geraniums (and many other plants) by name. The tissue culture propagation of morphologically conserved transgenes is not discussed.

U.S. Patent No. 5,286,635 to Hanson et al., entitled "Genetically, Transformed Pea Plants and Methods for Their Production," discloses the transfer of desired gene sequences into pea plants by incubating pea plant explants (preferably not callus) with Agrobacterium vectors containing the desired gene sequence. Mature seed material is used as the explant source. The issue of total morphologic conservation is not addressed.

**7.2.4.4 A method for commercially viable production of transgenic plants in which plants undergo minimal, and thus commercially acceptable, morphologic variation as a result of tissue culture propagation**

Thus while certain inroads have been made in the area of tissue culture plant propagation as well as in plant gene transfer, a need remains for a method for the commercially viable production of transgenic plants in which the plants undergo only minimal, and thus commercially acceptable, morphologic variation as a result of tissue culture propagation.

### 7.2.5 METHOD OF IDENTIFYING AND CHARACTERIZING THE ROLE OF INDIVIDUAL PLANT GENES IN QUANTITATIVE TRAIT EXPRESSION

When considering the application of biotechnology to plant improvement, a great deal of emphasis is usually placed on the strategy of introducing novel variability into plants via genetic engineering techniques. Over the past decade, advances have been made in developing methods of transferring genes to plant cells (see Potrykus et al., *Plant Mol. Bio. Rep.* 3:117-128 (1985)). For example, transfer and expression of single genes improving insect and herbicide resistance has reportedly been achieved in plants (Abel et al., *Science* 232:738-743 (1986); Shah et al., *Science* 233:478-481 (1986)). While there is excitement over advances in plant genetic engineering, the prospects for the general use of these techniques for plant improvement are tempered by the realization that very few genes corresponding to plant traits of interest have been identified or cloned.

One procedure that has been used by plant breeders to increase efficiency in the testing of traits which are difficult or expensive to evaluate is the use of indirect selection criteria (Hallaver and Miranda, *Quantitative Genetics in Corn Breeding* (Iowa State University Press 1981). One indirect selection criterion, for example, might be an easily recognized morphological characteristic of the plant which is either genetically linked to the desired trait or perhaps a component of the desired trait, e.g., the association between leaf size and seed size in beans.

Agronomically important traits such as, for example, plant yield, height, maturity, and fruit and grain characteristics, are all attractive targets for manipulation in plant improvement programs, but often have very low heritabilities. Heritability is the ratio of genetic to total variation and, therefore, is important to the efficiency of the selection process. Influencing heritability of such traits, sometimes termed "quantitative" traits, is difficult, however, because expression of a number of different gene products generally influences the phenotype. Quantitative traits are characterized by continuous rather than discrete distribution of phenotypic expression. There is currently a poor understanding of how single genes influence the expression of complex traits and, in conventional plant breeding programs, selection for inheritance of quantitative traits is difficult due to the unrecognized genetic basis of the trait. The use of direct gene transfer in manipulating these traits, of course, is therefore difficult due to problems in pinpointing and then

cloning those individual loci which contribute predominantly to the expression of the trait. Determination of genotypic information from phenotypic values is further imprecise because evaluation of the trait may frequently be confounded by environmental effects (Berger, "Multiple-Trait Selection Experiments: Current Status, Problem Areas and Experimental Approaches." In: Proceedings of the International Conference on Quantitative Genetics (Pollack et al., Eds.), p. 191-204 (Iowa St. Press 1977)).

Clearly, one area in which biotechnology may have a significant impact on plant improvement is in the development of new methods to identify and characterize the role of individual plant genes in quantitative trait expression. Following the development of a new class of plant molecular markers based on restriction fragment length polymorphisms, termed "RFLPs", (Helentjaris et al., Plant Mol. Bio. 5:109-118 (1985)) ("Helentjaris et al. I"), the processes to identify such loci and discriminate gene effects have been invented and are described and claimed herein. This and all other publications noted herein are hereby incorporated by reference. This will undoubtedly benefit plant improvement, not only within the context of conventional breeding approaches, but also by providing a means for identifying appropriate loci for future cloning and direct gene transfer efforts.

### **7.3. FUSION GENE INCLUDING A TRAIT AND A SILENCER DNA MOLECULE**

The present invention is directed to a DNA construct formed from a fusion gene which includes a trait DNA molecule and a silencer DNA molecule. The trait DNA molecule has a length that is insufficient to impart a desired trait to plants transformed with the trait DNA molecule. The silencer DNA molecule is operatively coupled to the trait DNA molecule with the trait and silencer DNA molecules collectively having sufficient length to impart the trait to plants transformed with the DNA construct. Expression systems, host cells, plants, and plant seeds containing the DNA construct are disclosed.

#### **7.3.1 FUSION GENE COMPRISING A PLURALITY OF TRAIT DNA MOLECULES**

In an alternative embodiment of the present invention, the DNA construct can be a fusion gene comprising a plurality of trait DNA molecules at least some of which having a length that is insufficient to impart that trait to plants transformed with that trait DNA molecule. However, the plurality of trait DNA molecules collectively have a length sufficient to impart their traits to plants transformed with the DNA construct and to effect post-transcriptional silencing of the fusion gene. Expression systems, host cells, plants, and plant seeds containing this embodiment of the DNA construct are disclosed.

#### **7.3.2 A Recombinant Chimeric Molecule**

The present invention also provides a recombinant chimeric DNA molecule comprising a plurality of DNA sequences each of which comprises a promoter operably linked to a DNA sequence which encodes a virus-associated protein, such as a coat protein (cp), a protease, or a replicase, wherein said DNA sequences are expressed in virus-susceptible plant cells transformed with said recombinant DNA molecule to impart resistance to infection by each of said viruses. Preferably, the DNA sequences are linked in tandem, i.e., exist in head to tail orientation relative to one another. Also, preferably substantially equal levels of resistance to infection by each of said viruses occurs in plant cells transformed with said plurality of DNA sequences.

Preferably, each DNA sequence is also linked to a 3' non- translated DNA sequence which functions in plant cells to cause the termination of transcription and the

addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequences. Preferably, the virus is a plant-associated virus, such as a potyvirus.

Thus, the present DNA molecule can be employed as a chimeric recombinant "expression construct," or "expression cassette" to prepare transgenic plants that exhibit increased resistance to infection by at least two plant viruses, such as potyviruses. The present cassettes also preferably comprise at least one selectable marker gene or reporter gene which is stably integrated into the genome of the transformed plant cells in association with the viral genes. The selectable marker and/or reporter genes facilitate identification of transformed plant cells and plants. Preferably, the virus gene array is flanked by two or more selectable marker genes, reporter genes or a combination thereof.

Another aspect of the present invention is a method of preparing a virus-resistant plant, such as a dicot, comprising:

(a) transforming plant cells with a chimeric recombinant DNA molecule comprising a plurality of DNA sequences, each comprising a promoter functional in said plant cells, operably linked to a DNA sequence, which encodes a protein associated with a virus which is capable of infecting said plant; (b) regenerating said plant cells to provide a differentiated plant; and (c) identifying a transformed plant which expresses the DNA sequences so as to render the plant resistant to infection by said viruses, preferably at substantially equal levels of resistance to infection by each virus.

Yet another object of the present invention is to provide a method for providing resistance to infection by viruses in a susceptible Cucurbitaceae plant which comprises:

(a) transforming Cucurbitaceae plant cells with a DNA molecule encoding a plurality of proteins from viruses which are capable of infecting said Cucurbitaceae plant; (b) regenerating said plant cells to provide a differentiated plant; and (c) selecting a transformed Cucurbitaceae which expresses the virus proteins at levels sufficient to render the plant resistant to infection by said viruses.

It is a further object of the present invention to provide multi-virus resistant transformed plant which contains stably-integrated DNA sequences encoding virus proteins.

### 7.3.3 CONTROLLING GENE EXPRESSION WITH EXTERNAL STIMULUS

The present invention involves, in one embodiment, the creation of a transgenic plant that contains a gene whose expression can be controlled by application of an external stimulus. This system achieves a positive control of gene expression by an external stimulus, without the need for continued application of the external stimulus to maintain gene expression. The present invention also involves, in a second embodiment, the creation of transgenic parental plants that are hybridized to produce a progeny plant expressing a gene not expressed in either parent. By controlling the expression of genes that affect the plant phenotype, it is possible to grow plants under one set of conditions or in one environment where one phenotype is advantageous, then either move the plant or plant its seed under another set of conditions or in another environment where a different phenotype is advantageous. This technique has particular utility in agricultural and horticultural applications.

In accordance with one embodiment of the invention, a series of sequences is introduced into a plant that includes a transiently-active promoter linked to a structural gene, the promoter and structural gene being separated by a blocking sequence that is in turn bounded on either side by specific excision sequences, a repressible promoter operably linked to a gene encoding a site-specific recombinase capable of recognizing the specific excision sequences, and a gene encoding a repressor specific for the repressible promoter whose function is sensitive to an external stimulus. Without application of the external stimulus, the structural gene is not expressed. Upon application of the stimulus, repressor function is inhibited, the recombinase is expressed and effects the removal of the blocking sequence at the specific excision sequences, thereby directly linking the structural gene and the transiently-active promoter.

In a modification of this embodiment, the sequences encoding the recombinase can be introduced separately into the plant via a viral vector.

In an alternative embodiment, no repressor gene or repressible promoter is used. Instead, the recombinase gene is linked to a germination-specific promoter and introduced into a separate plant from the other sequences. The plant containing the transiently-active promoter, blocking sequence, and structural gene is then hybridized with the plant containing the recombinase gene, producing progeny that contain all of the sequences.

When the second transiently-active promotor becomes active, the recombinase removes the blocking sequence in the progeny, allowing expression of the structural gene in the progeny, whereas it was not expressed in either parent.

In still another embodiment, the recombinase gene is simply linked to an inducible promoter. Exposure of the plant to the induce specific for the inducible promoter leads to the expression of the recombinase gene and the excision of the blocking sequence.

In all of these embodiments, the structural gene is expressed when the transiently-active promoter becomes active in the normal course of growth and development, and will continue to be expressed so long as the transiently-active promoter is active, without the necessity of continuous external stimulation. This system is particularly useful for developing seed, where a particular trait is only desired during the first generation of plants grown from that seed, or a trait is desired only in subsequent generations.

### 7.3.4 PREPARING PLANTS WHICH ARE RESISTANT TO MULTIPLE VIRUSES

It is still a further object of the present invention to provide virus resistant transformed plant cells which contain a plurality of viral genes, i.e., 2-7 or more genes, which are expressed as virus proteins, such as coat proteins, proteases and/or replicases, from the same virus strain, from different virus strains as from different members of the virus group, such as the potyvirus group. Representative viruses from which these DNA sequences can be isolated include, but are not limited to, potato virus X (PVX), potyviruses such as potato virus Y (PVY), cucumovirus (CMV), tobacco vein mottling virus, watermelon mosaic virus (WMV), zucchini yellow mosaic virus (ZYMV), bean common mosaic virus, bean yellow mosaic virus, soybean mosaic virus, peanut mottle virus, beet mosaic virus, wheat streak mosaic virus, maize dwarf mosaic virus, sorghum mosaic virus, sugarcane mosaic virus, johnsongrass mosaic virus, plum pox virus, tobacco etch virus, sweet potato feathery mottle virus, yam mosaic virus, and papaya ringspot virus (PRV), cucumoviruses, including CMA and comovirus.

Generally, a potyvirus is a single-stranded RNA virus that is surrounded by a repeating proteinaceous monomer, which is termed the coat protein (CP). The encapsidated virus has a flexuous rod morphology. The majority of the potyviruses are transmitted in a nonpersistent manner by aphids. As can be seen from the wide range of crops affected by potyviruses, the host range includes such diverse families of plants, but is not limited to Solanaceae, Chenopodiaceae, Gramineae, Compositae, Leguminosae, Dioscoreaceae, Cucurbitaceae, and Caricaceae.

The present invention is particularly directed to preparing plants which are resistant to multiple viruses. It is well known that particular plant types are often susceptible to more than one virus. Although PDR is an excellent approach to controlling the damaging effects of plant viruses, incorporating multiple virus resistance in a given plant can be challenging. For example, identifying and producing full length viral genes to transform plants can be expensive and time consuming. Further, such genes may be so large that they need to be incorporated in different expression systems which must be separately incorporated in plants.



#### **7.3.4.1 Using short fragments of viral genes to impart resistance**

Rather than attempting to incorporate full length viral genes in a plant, the present invention uses short fragments of such genes to impart resistance to the plant against a plurality of viral pathogens. These short fragments, which each by themselves have insufficient length to impart such resistance, are more easily and cost effectively produced than full length genes. There is no need to include in the plant separate promoters for each of the fragments; only a single promoter is required. Moreover, such viral gene fragments can preferably be incorporated in a single expression system to produce transgenic plants with a single transformation event.

##### **7.3.4.1.1 Example: papaya ringspot virus**

The impact of this simple strategy for multiple virus resistant transgenic plants could have far reaching effects in agriculture. An example is the case of papaya ringspot virus ("PRV"). Transgenic papaya with the coat protein gene of the PRV strain from Hawaii have been developed and found to be highly resistant under greenhouse and long-term field conditions. However, that papaya is largely susceptible to strains from other parts of the world, including Jamaica, Thailand, and Brazil.

Apparently, PRV resistance in papaya is highly specific and a number of transgenic papaya lines will need to be developed with different coat protein genes of the target countries to control the virus worldwide. With the present invention, a transgenic papaya could be developed with resistance to all PRV strains using viral gene fragments that total less than 1,000 base pairs plus a silencer DNA of about 400 bp; by comparison, the PRV coat protein gene alone is about 1,000 bp.

##### **7.3.4.1.2 Example: potato leaf roll virus, potato virus Y, and potato virus X**

Another use of the present invention involves imparting resistance against a plurality of different viruses. For example, in potato, the present invention can be employed to impart resistance against potato leaf roll virus, potato virus Y, and potato virus X. To effect such resistance, in accordance with the present invention, a DNA construct, driven by a single promoter, and containing a portion of the potato leafroll virus replicase encoding gene, a portion of the potato virus Y coat protein encoding gene, and a portion of the gene encoding the movement protein of potato virus X can be produced and

transformed into potato. As a result, transgenic potato with resistance to potato leafroll virus, potato virus Y, and potato virus X can be produced by a single transformation event. This constitutes a significant advance beyond incorporating full length versions of each of the genes with separate promoters together in a single expression vector or in separate vectors.

#### 7.3.4.1.3 Example: multiple resistance

Another use of the present invention involves imparting resistance to cucurbits against a number of viruses. For example, in squash, the present invention can be utilized to impart multiple resistance to zucchini yellow mosaic virus, papaya ringspot virus, watermelon mosaic virus IT, and squash mosaic virus. For example, a construct containing a portion of the coat protein encoding gene or a portion of the replicase encoding gene from each of these viruses, driven by a single promoter, can be produced and transformed into squash. The resulting transgenic squash is resistant to all of these viruses.

The present invention is exemplified primarily by the insertion of multiple virus cp expression cassettes into a binary plasmid and subsequent characterization of resulting plasmids. Combinations of CMV, ZYMV, WMV-2, SQMV, and PRV coat protein expression cassettes were placed in the binary plasmid pPRBN. Subsequently, binary plasmids harboring multiple cp expression cassettes were mobilized into Agrobacterium for use in plant transformation procedures. Binary plasmids harboring multiple expression cassettes are employed to transfer two or more virus coat protein transformation-susceptible genes into plants, such as members of the Cucurbitaceae family, along with the associated selectable marker and/or reporter genes.

### 7.3.5 IMPARTING OTHER TRAITS TO PLANTS

In addition to conferring on plants resistance to multiple viral diseases, the present invention can be utilized to impart other traits to plants. It is often desirable to incorporate a number of traits to a transgenic plant besides disease resistance. For example, color, enzyme production, etc. may be desirable traits to confer on a plant. However, transforming plants with a plurality of such traits encounter the same difficulties discussed above with respect to disease resistance. The present invention may be likewise useful in alleviating these problems with respect to traits other than disease resistance.

Thus, the present invention provides a genetic engineering methodology by which multiple traits can be manipulated and tracked as a single gene insert, i.e., as a construct which acts as a single gene which segregates as a single Mendelian locus. Although the invention is exemplified via virus resistance genes, in practice, any combination of genes could be linked. Therefore one could track a block of genes that provide traits such as disease resistance, plus enhanced herbicide resistance, plus extended shelf life, and the like, by simply tracking the linked selectable marker or reporter gene which has been incorporated into the transformation vector.

It was also discovered that when multiple tandem genes are inserted, they preferably all exhibit substantially the same degrees of efficacy, and more preferably substantially equal degrees of efficacy, wherein the term "substantial" as it relates to viral resistance is defined with reference to the assays described in the examples hereinbelow. For example, if one examines numerous transgenic lines containing an intact ZYMV and WMV-2 coat protein insert, one finds that if a line is immune to infection by ZYMV it is also immune to infection by WMV-2. Similarly, if a line exhibits a delay in symptom development to ZYMV it will also exhibit a delay in symptom development to WMV2. Finally, if a line is susceptible to ZYMV it will be susceptible to WMV-2. This phenomenon is unexpected. If there were not a correlation between the efficacy of each gene in these multiple gene constructs this approach as a tool in plant breeding would probably be prohibitively difficult to use. Even with single gene constructs, one must test numerous transgenic plant lines to find one that displays the appropriate level of efficacy. The probability of finding a line with useful levels of expression can range from 10-50% (depending on the species involved).

If the efficacy of individual genes in a Ti plasmid containing multiple genes were independent, the probability of finding a transgenic line that was resistant to each targeted virus would decrease dramatically. For example, in a species in which there is a 10% probability of identifying a line with resistance using a single gene insert, is transformed with a triple-gene construct CZW and each gene display an independent levels of efficacy, the probability of finding a line with resistance to CMV, ZYMV and WMV-2 would be  $0.1 \times 0.1 \times 0.1 = 0.001$  or 0.1 %. However, since the efficacy of multivalent genes is not independent of each other the probability of finding a line with resistance to CMV, ZYMV and WMV-2 is still 10% rather than 0.1 %. Obviously this advantage becomes more pronounced as constructs containing four or more genes are used.

### 7.3.6 PROBABILITY OF EXPRESSION

One problem with transforming plants to contain multiple traits is the possibility that not all of them will be successfully imparted. For example, where there are 4 new traits to be imparted to a transgenic plant, there is a 10% likelihood that each expression event will occur, making the probability of imparting all traits in a plant produced in accordance with the present invention much higher than in a plant transformed with full length trait genes driven by separate promoters. More particularly, the probability of expressing all 4 traits in the latter is 0.0001 (i.e.,  $0.1 \times 0.1 \times 0.1 \times 0.1$ ), while the probability in the present invention is 0.1.

### 7.3.7 COMMERCIALY VIABLE PRODUCTION METHOD

A need remains for a method for the commercially viable production of transgenic plants in which the plants undergo only minimal, and thus commercially acceptable, morphologic variation as a result of tissue culture propagation.

In order to meet this need, the present method is a process for commercially propagating plants by tissue culture in such a way as both to conserve desired plant morphology and to transform the plant with respect to one or more desired genes. The method includes the steps of (a) creating an Agrobacterium vector containing the gene sequence desired to be transferred to the propagated plant, preferably together with a marker gene; (b) taking one or more petiole explants from a mother plant and inoculating them with the Agrobacterium vector; (c) conducting callus formation in the petiole sections in culture, in the dark; and (d) culturing the resulting callus in growth medium having a benzylamino growth regulator such as benzylaminopurine or, most preferably, benzylaminopurine-riboside. Additional optional growth regulators including auxins and cytokinins (indole butyric acid, benzylamine, benzyladenine, benzylaminopurine, alpha naphthylacetic acid and others known in the art) may also be present.

Preferably, the petiole tissue is taken from *Pelargonium x domesticum* and the Agrobacterium vector contains an antisense gene for ACC synthase or ACC oxidase to prevent ACC synthase or ACC oxidase expression and, in turn, preventing ethylene formation. *Pelargoniums* propagated in culture using the present technique are resistant to wilting and petal shatter, and are morphologically conserved due to the use of petiole explants specifically and the particular culture media disclosed.

### 7.3.8 PRODUCTION OF A MATURE HETEROLOGOUS PROTEIN IN TRANSFORMED MONOCOT PLANT CELLS

In one aspect, the invention includes a method of producing, in monocot plant cells, a mature heterologous protein selected from the group consisting of (i) mature, glycosylated  $\alpha_1$ -antitrypsin (AAT) having the same N-terminal amino acid sequence as mature AAT produced in humans and a glycosylation pattern which increases serum half-life substantially over that of non-glycosylated mature AAT; (ii) mature, glycosylated antithrombin III (ATIII) having the same N-terminal amino acid sequence as mature ATIII produced in humans; (iii) mature human serum albumin (HSA) having the same N-terminal amino acid sequence as mature HSA produced in humans and having the folding pattern of native mature HSA as evidenced by its bilirubin-binding characteristics; and (iv) mature, active subtilisin BPN' (BPN'), glycosylated or non-glycosylated, having the same N-terminal amino acid sequence as BPN' produced in *Bacillus*.

The method includes obtaining monocot cells transformed with a chimeric gene having (i) a monocot transcriptional regulatory region, inducible by addition or removal of a small molecule, or during seed maturation, (ii) a first DNA sequence encoding the heterologous protein, and (iii) a second DNA sequence encoding a signal peptide. The second DNA sequence is operably linked to the transcriptional regulatory region and to the first DNA sequence. The first DNA sequence is in translation-frame with the second DNA sequence, and the two sequences encode a fusion protein.

The transformed cells are cultivated under conditions effective to induce the transcriptional regulatory region, thereby promoting expression of the fusion protein and secretion of the mature heterologous protein from the transformed cells. The mature heterologous protein produced by the transformed cells is then isolated.

In one embodiment of the method, the first DNA sequence encodes pro-subtilisin BPN' (proBPN'), the cultivating includes cultivating the transformed cells at a pH between 5 and 6, and the isolating step includes incubating the proBPN' to under condition effective to allow its autoconversion to active mature BPN'. In another embodiment, the first DNA sequence encodes mature BPN', and the cells are transformed with a second chimeric gene containing (i) a transcriptional regulatory region inducible by addition or

removal of a small molecule, (ii) a third DNA sequence encoding the pro-peptide moiety of BPN', and (iii) a fourth DNA sequence encoding a signal polypeptide. The fourth DNA sequence is operably linked to the transcriptional regulatory region and to the third DNA sequence, and the signal polypeptide is in translation-frame with the pro-peptide moiety and is effective to facilitate secretion of expressed pro-peptide moiety from the transformed cells. The cultivating step includes cultivating the transformed cells at a pH between 5 and 6, and the isolating step includes incubating the mature BPN and the pro-moiety under conditions effective to allow the conversion of BPN' by the pro-moiety to active mature BPN'.

#### **7.3.8.1 Inducing the transcriptional regulatory region**

In other embodiments of the method, the transcriptional regulatory region may be a promoter derived from a rice or barley  $\alpha$ -amylase gene, including RAmylA, RAmylB, RAmy2A, RAmy3A, RAmy3B, RAmy3C, RAmy3D, RAmy3E, pM/C, gKAmyl41, gKAmyl55, Amy32b, or HV18. The chimeric gene may further include, between the transcriptional regulatory region and the fusion protein coding sequence, the 5' untranslated region (5' UTR) of an inducible monocot gene such as one of the rice or barley  $\alpha$ -amylase genes described above. One preferred 5' UTR is that from the RAmylA gene, which is effective to enhance the stability of the gene transcript. The chimeric gene may further include, downstream of the coding sequence, the 3' untranslated region (3' UTR) from an inducible monocot gene, such as one of the rice or barley  $\alpha$ -amylase genes mentioned above. One preferred 3'UTR is from the RAmylA gene.

#### **7.3.8.2 Preferred Promoters**

Where the method is employed in protein production in a monocot cell culture, preferred promoters are the RAmy3D and RAmy3E gene promoters, which are upregulated by sugar depletion in cell culture. Where the gene is employed in protein production in germinating seeds, a preferred promoter is the RAmylA gene promoter, which is upregulated by gibberellic acid during seed germination. Where gene is upregulated during seed maturation, a preferred promoter is the barley endosperm-specific BI-hordein promoter.

#### **7.3.8.3 Product: A Mature Heterologous Protein**



The invention also includes a mature heterologous protein produced by the above method. The protein has a glycosylation pattern characteristic of the monocot plant in which the protein is produced. The glycosylated protein is selected from the group consisting of (i) mature glycosylated  $\alpha_1$ -antitrypsin (AAT) having the same N-terminal amino acid sequence as mature AAT produced in humans and having a glycosylation pattern which increases serum half-life substantially over that of non-glycosylated mature AAT; (ii) mature glycosylated antithrombin III (ATIII) having the same N-terminal amino acid sequence as mature ATIII produced in humans; and (iii) mature glycosylated subtilisin BPN' (BPN') having the same N-terminal amino acid sequence as BPN' produced in *Bacillus*.

The invention also includes plant cells and seeds capable of producing the mature heterologous proteins according to the above method.

These and other objects and features of the invention will be more fully understood when the following detailed description of the invention is read in conjunction with the accompanying drawings.

### **7.3.9 IDENTIFYING SUCH LOCI AND DISCRINATING GENE EFFECTS OF RESTRICTION FRAGMENT LENGTH POLYMORPHISMS**

The development of new methods to identify and characterize the role of individual plant genes in quantitative trait expression has significant impact on plant improvement. Following the development of a new class of plant molecular markers based on restriction fragment length polymorphisms, termed "RFLPs", (Helentjaris et al., Plant Mol. Bio. 5:109-118 (1985)) ("Helentjaris et al. I"), the processes to identify such loci and discriminate gene effects have been invented and are described and claimed herein.

RFLPs are differences observed between genotypes in the fragment lengths of restriction endonuclease-digested DNA. RFLPs occur as a result of base pair or positional changes in the restriction enzyme recognition sites which flank a chromosomal location and can be detected by hybridization of labelled DNA clones containing sequences that are homologous to a portion of the chromosomal fragment. Hybridization with a unique cloned sequence can permit the identification of a specific chromosomal region (locus).

This technology employs cloned DNA fragments to detect differences between individuals at the DNA sequence level. When genomic DNAs from two genetically distinct individuals are digested with a restriction enzyme, electrophoresed and probed with a labelled DNA clone, polymorphisms in the hybridization patterns sometimes result due to sequence differences between the individuals. The term "restriction fragment length polymorphism" has been coined to describe this variation.

#### **7.3.9.1 Using RFLPs as Genetic Markers**

Differences in fragment lengths which are revealed, for example, by agarose gel electrophoresis, function as alleles of that RFLP. Thus, RFLPs can serve as genetic markers in a manner analogous to conventional morphological or isozyme markers. Unlike most genetic markers, however, they are not the products of transcription and translation. Additionally, RFLP possess certain additional advantages over previously available genetic markers. First, RFLPs reflect existing differences between genetically distinct individuals. The potential number of RFLPs for all practical purposes is thus unlimited, as digestion of the genomic DNA of any higher eukaryote with a six base recognition enzyme will generate more than a million fragments, many of which can be polymorphic.

Additionally, over one hundred different restriction enzymes have now been described, each of which may generate a new and different set of fragments (Roberts, *Nuc. Acids Res.* 10:117-144 (1982)). The utility of isozyme markers or morphological markers in studies is frequently limited by a lack of informativeness in lines of interest or by an insufficient availability or chromosomal distribution of the loci.

#### 7.3.9.2 Isozyme Variation in Plant Breeding

The use of isozyme variation in plant breeding is, like RFLP technology, one of indirect selection. (Tanksley and Orton, *Isozymes in Plant Genetics and Breeding 1B* (Elsevier, N.Y. 1983). The time required to backcross a trait from a donor to a recurrent parent is the product of the generation time by the number of generations. Therefore, screening for traits linked to isozymes, which may sometimes be identified in seeds or seedlings, can reduce the time required for evaluation, especially if the expression of that trait is controlled by recessive alleles. In addition, the number of backcross generations necessary to sufficiently recover the phenotype of the recurrent parent can be reduced by selection for isozymes associated with the recurrent parent.

In tomato, closely linked isozyme variation has been used to follow the inheritance of several simply inherited traits such as, for example, nematode resistance (Rick and Fobes, *Rep. Tomato Genet. Coop.* 24:25 (1974)). The inheritance of quantitative traits has also been followed by the use of multiple isozyme loci. For example, in tomato, eleven isozyme loci were used to survey eight of the 12 chromosomes and three independent genetic factors were detected in association with cold tolerance (Vallejos and Tanksley, *Theor. Appl. Genet.* 66:241-247 (1983)). In maize, changes in the frequency of eight isozyme loci were found to be associated with selection for improved grain yield (Stuber et al., *Crop. Sci.* 22:737-740 (1982)). Nevertheless, for a multigenic trait, only that portion of the genome closely linked to an isozyme marker can be considered, and many other major or minor genes may be associated with the expression of quantitative traits.

Maize is perhaps the best characterized plant system in terms of isozymes and yet only about two dozen isozyme loci have been located and it is rare for more than a dozen of these to be informative in any particular cross involving Corn Belt germplasm. By contrast, using the inventors' RFLP technology, over 300 RFLPs covering all ten maize chromosomes have been characterized (Helentjaris et al., *Trends in Genetics.* 3:217-221

(1987)). The level of informativeness of these RFLPs is great. In a study involving the maize cross Tx303X Co159, only 13 informative isozyme loci with very biased coverage of less than the 10 chromosomes were available. In contrast, using RFLP analysis, more than 99 informative RFLPs covering all 10 chromosomes can be analyzed in the same cross (unpublished data). In a recent comparative survey of Corn Belt germplasm RFLPs averaged greater than five alleles per locus while isozyme loci averaged less than two (M. Walton and T. Helentjaris, abstract).

### **7.3.9.3 RFLP Markers in Plant Breeding**

RFLP markers rarely possess detectable phenotype effects of their own, so they can be utilized in economic lines without detriment and many can be evaluated at one time without the pleiotropic effects often seen with phenotypic markers. Evaluation can be performed on small amounts of DNA obtained from plant tissue at virtually any stage of plant development from roots, to shoots, to fruits, or even with tissue culture material. Evaluation of RFLPs is not affected by environmental factors and greenhouse-grown plants will not differ from field-grown plants when tested. Finally, the evaluation of RFLPs reveals the exact genotype, so the heterozygous state can be differentiated from the homozygous condition at any chromosomal location.

Many of the potential applications and theoretical advantages of RFLPs compared to more conventional phenotypic or isozyme marking systems have been described previously. Helentjaris et al. I. In one application of the use of RFLP markers in plant studies, genetic linkage maps based on these markers have been constructed for both maize and tomato (Helentjaris et al., *Theor. Appl. Genet.* 72:761-769 (1986)) ("Helentjaris et al. II"). Similar linkage maps are also being constructed for other crop species, such as Brassica Figdore et al., *Theor. Appl. Genetics.* 75:833-840 (1988); Slocum et al., In "Genetic Maps" (S. J. O'Brien, ed.), 5th Edition, Cold Spring Harbor Press, N.Y. (1990)). Close to 120 RFLPs in tomato have been arranged into linkage groups by comparing segregation patterns in an F2 population derived from homozygous parental lines. Approximately 70 RFLPs have been mapped by another group of workers. Bernatzky and Tanksley, *Genetics* 112:887- 898 (1986). Over 300 RFLPs in maize have been arranged into linkage groups (Helentjaris et al., 1986; Helentjaris et al., 1987). The locations of the maize RFLP loci have been correlated to the conventional maize genetic map by analyzing

the inheritance patterns of the RFLPs in maize lines monosomic for different chromosomes (Helentjaris et al., Proc. Nat. Acad. Sci. USA 83:6035-6039 (1986)) ("Helentjaris et al. III"), by establishing linkage relationships with isozyme markers, cloned genes, and morphological markers with previously identified chromosomal locations (Wright et al., MNL 61:89-90 (1987)), and by analyzing inheritance patterns in B-A translocation stocks (Helentjaris et al., Weber and Helentjaris, Genetics 121:583-590 (1989). The resulting map shows the resolution possible.

Numerous direct applications of RFLP technology to facilitate plant breeding programs have been suggested. Helentjaris et al. I. Because of the large numbers of RFLP markers available in a population of interest, one of the more important applications of RFLPs may be as markers linked to genes affecting the expression of quantitatively inherited traits. In this application, RFLPs function as indirect selection criteria for traits which are difficult or expensive to evaluate phenotypically. A prerequisite for the use of RFLPs as indirect selection criteria is the identification of RFLPs closely linked to the quantitative trait loci (QTL) affecting expression of the trait of interest.

Currently, the introgression of quantitative traits from one germplasm to another involves the identification of favorable genotypes in segregating generations followed by repeated backcrossing to commercially acceptable cultivars. This procedure is feasible for simply inherited quantitative traits, but as the number of genes controlling a trait increases, screening the number of F<sub>2</sub> segregants required to identify at least one individual which represents the ideal (homozygous) genotype quickly becomes prohibitive. For example, with one gene and two alleles of equal frequency, the probability of recovering a desirable genotype in the F<sub>2</sub> generation is 1/4. However, if the number of genes is increased to 5 or 10, the probability of recovering an ideal genotype in the F<sub>2</sub> population is reduced to approximately one in one thousand and one in one million, respectively. Thus, to identify desirable segregants, one must either reduce the number of segregants needed or have available very efficient screening procedures. Additionally, in situations where environmental effects interfere with the ability to draw accurate genotypic information from the phenotype, large allocations of time and resources are required to evaluate progeny in replicated trials within several target environments.

Described and claimed herein is the use of RFLPs to dissect multigenic traits into their individual genetic components. A genome, or portion thereof, saturated with RFLPs or probed with select RFLP markers, all of which can be evaluated together in individual plants, has been found to give the resolution necessary to break down traits of complex inheritance into individual loci, even those under a significant environmental influence. The procedure is equally workable with dominant or recessive traits and can be used to accelerate introgression of desired genes into a commercially acceptable cultivar. As used herein, "plants" includes all forms of plant life, such as crop plants, mushrooms and fungi, ferns, trees, flowers and so on. These examples are not intended to be limiting but are merely illustrative of the wide applications of the invention.

## PARTICULAR DEFINITIONS

The terms below have the following meaning, unless indicated otherwise in the specification.

As used in this specification, a "transiently-active promoter" is any promoter that is active either during a particular phase of plant development or under particular environmental conditions, and is essentially inactive at other times.

A "plant active promoter" is any promoter that is active in cells of a plant of interest. Plant-active promoters can be of viral, bacterial, fungal, animal or plant origin.

A gene that results in an altered plant phenotype is any gene whose expression leads to the plant exhibiting a trait or traits that would distinguish it from a plant of the same species not expressing the gene. Examples of such altered phenotypes include a different growth habit, altered flower or fruit color or quality, premature or late flowering, increased or decreased yield, sterility, mortality, disease susceptibility, altered production of secondary metabolites, or an altered crop quality such as taste or appearance.

A gene and a promoter are to be considered to be operably linked if they are on the same strand of DNA, in the same orientation, and are located relative to one another such that the promoter directs transcription of the gene (i.e. in cis). The presence of intervening DNA sequences between the promoter and the gene does not preclude an operable relationship.

A "blocking sequence" is a DNA sequence of any length that blocks a promoter from effecting expression of a targeted gene.

A "specific excision sequence" is a DNA sequence that is recognized by a site-specific recombinase.

A "recombinase" is an enzyme that recognizes a specific excision sequence or set of specific excision sequences and effects the removal of, or otherwise alters, DNA between specific excision sequences.

A "repressor element" is a gene product that acts to prevent expression of an otherwise expressible gene. A repressor element can comprise protein, RNA or DNA.

A "repressible promoter" is a promoter that is affected by a repressor element, such that transcription of the gene linked to the repressible promoter is prevented.

"Expression" means transcription or transcription followed by translation of a particular DNA molecule.

As used herein, with respect to a DNA sequence or "gene", the term "isolated" is defined to mean that the sequence is either extracted from its context in the viral genome by chemical means and purified and/or modified to the extent that it can be introduced into the present vectors in the appropriate orientation, i.e., sense or antisense.

"Cell culture" refers to cells and cell clusters, typically callus cells, growing on or suspended in a suitable growth medium.

"Germination" refers to the breaking of dormancy in a seed and the resumption of metabolic activity in the seed, including the production of enzymes effective to break down starches in the seed endosperm.

"Inducible" means a promoter that is upregulated by the presence or absence of a small molecules. It includes both indirect and direct inducement.

"Inducible during germination" refers to promoters which are substantially silent but not totally silent prior to germination but are turned on substantially (greater than 25%) during germination and development in the seed. Examples of promoters that are inducible during germination are presented below.

"Small molecules", in the context of promoter induction, are typically small organic or bioorganic molecules less than about 1 kDal. Examples of such small molecules



include sugars, sugar-derivatives (including phosphate derivatives), and plant hormones (such as, gibberellic or abscisic acid).

"Specifically regulatable" refers to the ability of a small molecule to preferentially affect transcription from one promoter or group of promoters (e.g., the  $\alpha$ -amylase gene family), as opposed to non-specific effects, such as, enhancement or reduction of global transcription within a cell by a small molecule.

"Seed maturation" or "grain development" refers to the period starting with fertilization in which metabolizable reserves, e.g., sugars, oligosaccharides, starch, phenolics, amino acids, and proteins, are deposited, with and without vacuole targeting, to various tissues in the seed (grain), e.g., endosperm, testa, aleurone layer, and scutellar epithelium, leading to grain enlargement, grain filling, and ending with grain desiccation.

"Inducible during seed maturation" refers to promoters which are turned on substantially (greater than 25%) during seed maturation.

"Heterologous" is defined to mean not identical, e.g. different in nucleotide and/or amino acid sequence, phenotype or an independent isolate.

"Heterologous DNA" or "foreign DNA" refers to DNA which has been introduced into plant cells from another source, or which is from a plant source, including the same plant source, but which is under the control of a promoter or terminator that does not normally regulate expression of the heterologous DNA.

"Heterologous protein" is a protein, including a polypeptide, encoded by a heterologous DNA. A "transcription regulatory region" or "promoter" refers to nucleic acid sequences that influence and/or promote initiation of transcription. Promoters are typically considered to include regulatory regions, such as enhancer or inducer elements.

"Chimeric" is defined to mean the linkage of two or more DNA sequences which are derived from different sources, strains or species, i.e., from bacteria and plants, or that two or more DNA sequences from the same species are linked in a way that does not occur

in the native genome. Thus, the DNA sequences useful in the present invention may be naturally occurring, semi-synthetic or entirely synthetic. The DNA sequence may be linear or circular, i.e., may be located on an intact or linearized plasmid, such as the binary plasmids described below.

A "chimeric gene," in the context of the present invention, typically comprises a promoter sequence operably linked to DNA sequence that encodes a heterologous gene product, e.g., a selectable marker gene or a fusion protein gene. A chimeric gene may also contain further transcription regulatory elements, such as transcription termination signals, as well as translation regulatory signals, such as, termination codons.

"Operably linked" refers to components of a chimeric gene or an expression cassette that function as a unit to express a heterologous protein. For example, a promoter operably linked to a heterologous DNA, which encodes a protein, promotes the production of functional mRNA corresponding to the heterologous DNA.

A "product" encoded by a DNA molecule includes, for example, RNA molecules and polypeptides.

"Removal" in the context of a metabolite includes both physical removal as by washing and the depletion of the metabolite through the absorption and metabolizing of the metabolite by the cells.

"Substantially isolated" is used in several contexts and typically refers to the at least partial purification of a protein or polypeptide away from unrelated or contaminating components.

Methods and procedures for the isolation or purification of proteins or polypeptides are known in the art.

"Stably transformed" as used herein refers to a cereal cell or plant that has foreign nucleic acid stably integrated into its genome which is transmitted through multiple generations.

" $\alpha_1$ -antitrypsin or "AAT" refers to the protease inhibitor which has an amino acid sequence substantially identical or homologous to AAT protein.

"Antithrombin III" or "ATIII" refers to the heparin-activated inhibitor of thrombin and factor Xa, and which has an amino acid sequence substantially identical or homologous to ATIII protein.

Human serum albumin" or "HSA" refers to a protein which has an amino acid sequena substantially identical or homologous to the mature HSA protein.

"Subtilisin" or "subtilisin BPN" or "BPN" refers to the protease enzyme produced naturally by *B. amyloliquefaciens*.

"proBPN" refers to a form of BPN' having an approximately 78 amino-acid "pro" moiety that functions as a chaperon polypeptide to assist in folding and activation of the BPN'.

"Codon optimization" refers to changes in the coding sequence of a gene to replace native codons with those corresponding to optimal codons in the host plant.

A DNA sequence is "derived from" a gene, such as a rice or barley  $\alpha$ -amylase gene, if it corresponds in sequence to a segment or region of that gene. Segments of genes which may be derived from a gene include the promoter region, the 5' untranslated region, and the 3' untranslated region of the gene.

#### 7.4.1 GENERAL APPROACH

Generally, the nomenclature and laboratory procedures with respect to standard recombinant DNA technology can be found in Sambrook, et al., MOLECULAR - CLONING - A LABORATORY MANUAL, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1989 and in S.B. Gelvin and R.A. Schilperoort, PLANT MOLECULAR BIOLOGY, 1988. Other general references are provided throughout this document. The procedures therein are known in the art and are provided for the convenience of the reader.

Most of the recombinant DNA methods employed in practicing the present invention are standard procedures, well known to those skilled in the art, and described in detail in, or example, European Patent Application Publication Number 223,452, published November 29, 1986, which is incorporated herein by reference. Enzymes are obtained from commercial sources and are used according to the vendor's recommendations or other variations known in the art. General references containing such standard techniques include the following: R. Wu, ed. (1979) Methods Emzymology, Vol. 68; J.H. Miller (1972) Experiments in Molecular Genetics; J. Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual 2nd Ed.; D.M. Glover, ed. (1985) DNA Cloning Vol. II; H.G. Polites and K.R. Marotti (1987) "A step-wise protocol for cDNA synthesis, " Biotechniques 4; 514-520; S.B. Gelvin and R.A. Schilperoort, eds. Introduction, Expression, and Analysis of Gene Products in Plants, all of which are incorporated by reference.

## **7.4.2 CONTROL OF PLANT GENE EXPRESSION**

### **7.4.2.1 Using a Transiently-active Promoter**

This invention relates to a method of creating transgenic plants wherein the expression of certain plant traits is ultimately under external control. In one embodiment the control is achieved through application of an external stimulus; in another embodiment it is achieved through hybridization, in still another embodiment it is achieved by direct introduction of a recombinase or recombinase gene into a plant. The transgenic plants of the present invention are prepared by introducing into their genome a series of functionally interrelated DNA sequences, containing the following basic elements: a plant-active promoter that is active at a particular stage in plant development or under particular environmental conditions ("transiently- active promoter"), a gene whose expression results in an altered plant phenotype which is linked to the transiently-active promoter through a blocking sequence separating the transiently-active promoter and the gene, unique specific excision sequences flanking the blocking sequence, wherein the specific excision sequences are recognizable by a site- specific recombinase, a gene encoding the site-specific recombinase, an alternative repressible promoter linked to the recombinase gene, and an alternative gene that encodes the repressor specific for the repressible promoter, the action of the repressor being responsive to an applied or exogenous stimulus. While these elements may be arranged in any order that achieves the interactions described below, in one embodiment they are advantageously arranged as follows: a first DNA sequence contains the transiently-active promoter, a first specific excision sequence, the blocking sequence, a second specific excision sequence, and the gene whose expression results in an altered plant phenotype; a second DNA sequence contains the repressible promoter operably linked to the recombinase gene, and optionally an enhancer; and a third DNA sequence containing the gene encoding the repressor specific for the repressible promoter, itself linked to a promoter functional and constitutive in plants. The third DNA sequence can conveniently act as the blocking sequence located in the first DNA sequence, but can also occur separately without altering the function of the system. This embodiment can be modified such that the recombinase sequence is introduced separately via a viral vector. In an alternative embodiment, an advantageous arrangement is as follows: a first plant containing a DNA sequence comprising the transiently-active promoter, a first specific excision signal sequence, the blocking sequence, a second specific excision signal sequence, and the gene whose expression results in an altered plant phenotype; a second

plant containing a DNA sequence comprising a constitutive plant-active promotor operably linked to the recombinase gene (the two plants being hybridized to produce progeny that contain all of the above sequences).

When a plant contains the basic elements of either embodiment, the gene whose expression results in an altered plant phenotype is not active, as it is separated from its promoter by the blocking sequence. In the first embodiment, absent the external stimulus, the repressor is active and represses the promoter that controls expression of the recombinase; in the alternative embodiment the recombinase is not present in the same plant as the first DNA sequence. Such a plant will not display the altered phenotype, and will produce seed that would give rise to plants that also do not display the altered phenotype. When the stimulus to which the repressor is sensitive is applied to this seed or this plant, the repressor no longer functions, permitting the expression of the site-specific recombinase, or alternatively, when the recombinase is introduced via hybridization it is expressed during germination of the seed, either of which effects the removal of the blocking sequence between the specific excision signal sequences. Upon removal of the blocking sequence, the transiently-active promoter becomes directly linked to the gene whose expression results in an altered plant phenotype. A plant grown from either treated or hybrid seed, or a treated plant, will still not exhibit the altered phenotype, until the transiently-active promoter becomes active during the plant's development, after which the gene to which it is linked is expressed, and the plant will exhibit an altered phenotype.

#### **7.4.2.2 Transgenic Plants that Produce Seeds that Cannot Germinate**

In a preferred embodiment, the present invention involves a transgenic plant or seed which, upon treatment with an external stimulus produces plants that produce seed that cannot germinate (but that is unaltered in other respects). If the transiently-active promoter is one that is active only in late embryogenesis, the gene to which it is linked will be expressed only in the last stages of seed development or maturation. If the gene linked to this promoter is a lethal gene, it will render the seed produced by the plants incapable of germination. In the initially-transformed plant cells, this lethal gene is not expressed, not only because the promoter is intrinsically inactive, but because there is a blocking sequence separating the lethal gene from its promoter. Also within the genome of

these cells are the genes for the recombinase, linked to a repressible promoter, and the gene coding for the repressor. The repressor is expressed constitutively and represses the expression of the recombinase. These plant cells can be regenerated into a whole plant and allowed to produce seed. The mature seed is exposed to a stimulus, such as a chemical agent, that inhibits the function of the repressor. Upon inhibition of the repressor, the promoter driving the recombinase gene is depressed and the recombinase gene is expressed. The resulting recombinase recognizes the specific excision sequences flanking the blocking sequence, and effects the removal of the blocking sequence. The late embryogenesis promoter and the lethal gene are then directly linked. The lethal gene is not expressed, however, because the promoter is not active at this time in the plant's life cycle. This seed can be planted, and grown to produce a desired crop of plants. As the crop matures and produces a second generation of seed, the late embryogenesis promoter becomes active, the lethal gene is expressed in the maturing second generation seed, which is rendered incapable of germination. In this way, accidental reseeding, escape of the crop plant to areas outside the area of cultivation, or germination of stored seed can be avoided.

#### **7.4.2.3 Transgenic Plants Hybridized to Display Phenotype Not Seen in Either Parent**

In an alternative preferred embodiment, the present invention involves a pair of transgenic plants that are hybridized to produce progeny that display a phenotype not seen in either parent. In this alternative embodiment a transiently-active promoter that is active only in late embryogenesis can be linked to a lethal gene, with an intervening blocking sequence bounded by the specific excision sequences. These genetic sequences can be introduced into plant cells to produce one transgenic parent plant. The recombinase gene is linked to a germination-specific promoter and introduced into separate plant cells to produce a second transgenic parent plant. Both of these plants can produce viable seed if pollinated. If the first and second transgenic parent plants are hybridized, the progeny will contain both the blocked lethal gene and the recombinase gene. The recombinase is expressed upon germination of the seed and effects the removal of the blocking sequence, as in the first embodiment, thereby directly linking the lethal gene and the transiently-active promoter. As in the first embodiment, this promoter becomes active during maturation of the second generation seed, resulting in seed that is incapable of germination. Ideally, the first parent employs a male-sterility gene as the blocking sequence, and includes an herbicide resistance gene. In this way, self-pollination of the

first transgenic parent plant is avoided, and self-pollinated second transgenic parent plants can be eliminated by application of the herbicide. In the hybrid progeny, the male-sterility gene is removed by the recombinase, resulting in hybrid progeny capable of self-pollination.

#### **7.4.2.3.1 Linking the Recombinase Gene to an Inducible Promoter**

In another embodiment, the recombinase gene is linked to an inducible promoter. Examples of such promoters include the copper, controllable gene expression system (Mett et al., 1993) and the steroid- inducible gene system (Schena et al., 1991). Exposure of the transgenic plant to the inducer specific for the inducible promoter leads to expression of the recombinase gene and the excision of the blocking sequence. The gene that results in an altered plant phenotype is then expressed when the transiently active promoter becomes active.

#### **7.4.2.3.2 Selection of an Appropriate Promoter**

Any appropriate transiently-active promoter can be used, and selection of an appropriate promoter will be governed by such considerations as plant type and the phenotypic trait over which control is sought. The transiently-active promoter is preferably not a "leaky" promoter, meaning that it is active substantially only during a well- defined phase of plant growth or under particular environmental conditions, and substantially inactive at all other times. This property prevents the premature "triggering" of the system. There are numerous published examples of transiently-active promoters, which can be applied in the present system. The principle consideration for selecting an appropriate promoter is the stage in the plant's life at which it is desired to have the altered phenotype expressed. If it is desired to have the phenotype expressed after the first generation, a promoter that is active during seed production is preferred, as it will not be active during the vegetative phase of first generation plant growth. If it is desired to have the altered phenotype expressed at some time during the first generation itself, a promoter that is active at an earlier stage would be appropriate. It will be readily apparent to workers conversant in the art that the timing of the application of the external stimulus to the plant to trigger the system, in those embodiments employing the repressible promotor system, should occur prior to the stage at which the selected transiently-active promotor is active for the generation of plant which is desired to display the altered phenotype. A promoter



active in late embryogenesis, such as the LEA promoter, Hughes and Galau, 1989 and 1991, Galau, et al., 1991, 1992 and 1993, is ideal when it is desired to have the altered phenotype appear after the first generation, because it is active only during the formation of the embryo within the seed, after the first generation plant has completed a season of vegetative growth (embryogenesis is virtually the last stage in seed formation, after most other fruit and seed structures are formed).

#### **7.4.2.3.3 Gene(s) Linked to the Plant Development Promoter**

The gene or genes linked to the plant development promoter can be any gene or genes whose expression results in a desired detectable phenotype. This phenotype could be any trait that would be desired in a plant in one situation, but not desired in another, such as male sterility, drought resistance, insect resistance, early or late seed germination, or early or late flowering, to give a few examples. Often a plant can possess traits that are advantageous in some ways or under some conditions, but at a certain cost to the plant. For instance, a trait for insect resistance might involve the production of secondary plant metabolites or structures, at a certain metabolic expense to the plant. This is advantageous in an environment where pests are present, but essentially an unnecessary-burden where they are not. Another example is the production of seeds in an annual fruit crop, such as watermelon. Obviously, it is necessary for at least one generation of plants to produce seeds, so that a seed company can produce seed for sale to growers, but a seedless fruit crop grown from that seed is commercially desirable. Still another example is a trait that allows ready and rapid seed germination in a cereal crop. This is advantageous for getting a crop established as rapidly as possible and with a minimum of effort, but very undesirable if it leads to germination of the harvested grain in the grain bin. Still another example would be where the plant is desirable in one location or season (as a winter forage crop, for instance), but considered a weed in another. If the second generation seed were incapable of germination, it would prevent post-harvest germination, the "tescape" of a plant through natural seed dispersal into a location where it is not desired, or accidental reseeding. These last two examples could advantageously employ a lethal gene (meaning a gene whose expression somehow interferes in plant growth or development), so that the second generation seed simply will not germinate, or the last example could alternatively employ any gene that introduces a trait that decreases the plant's vigor, such as disease susceptibility, early flowering, low seed production, or seedless. A ribosomal

inhibitor protein ("RIP") gene is a preferred lethal gene, the saponin 6 RIP, (GenBank ID SOSAPG, Accession No. X15655), being particularly preferred. RIP directly interferes in the expression of all protein in a plant cell, without being toxic to other organisms. Expression of RIP in the cells of the embryo would entirely prevent germination of the seed.

#### 7.4.2.3.4 Blocking Sequence

The blocking sequence can be any sequence that prevents expression of the gene linked to the transiently-active promotor, such as a termination signal, but in those embodiments employing a repressible promoter is advantageously the sequence that codes for the repressor. In this way, when the blocking sequence is excised, the repressor gene is eliminated, thus further minimizing the chance of later inhibition of the system. In the hybrid embodiment, the blocking sequence is advantageously a gene that produces male sterility (such as a lethal gene linked to an anther-specific promotor). In this way, hybridization is facilitated, but hybrid progeny will be capable of self-pollination when the blocking sequence is removed by the recombinase.

#### 7.4.2.3.5 Repressible Promoter System

In those embodiments employing a repressible promoter system, the gene encoding the repressor is responsive to an outside stimulus, or encodes a repressor element that is itself responsive to an outside stimulus, so that repressor function can be controlled by the outside stimulus. The stimulus is preferably one to which the plant is not normally exposed, such as a particular chemical, temperature shock, or osmotic shock. In this way, the simple application of the stimulus will block the repression of the recombinase, yet there will be a low probability of the repressor being accidentally or incidentally blocked. If the repressor is sensitive to a chemical stimulus, the chemical is preferably non-toxic to the crop and to non-pest animals. A preferred system is the Tn10 tet repressor system, which is responsive to tetracycline. Gatz and Quail (1988); Gatz, et al. (1992). In this system, a modified Cauliflower Mosaic Virus (CaMV) 35S promoter containing one or more, preferably three, tet operons is used; the Tn10 tet repressor gene produces a repressor protein that binds to the tet operon(s) and prevents the expression of the gene to which the promoter is linked. The presence of tetracycline inhibits binding of the Tn10 tet repressor to the tet operon(s), allowing free expression of the linked gene. This system is

preferred because the stimulus, tetracycline, is not one to which the plant would normally be exposed, so its application can be controlled. Also, since tetracycline has no harmful effects on plants or animals, its presence would not otherwise impede the normal development of the plant, and residual amounts left on the seed or plant after treatment would have no significant environmental impact. Examples of other repressible promoter systems are described by Lanzer and Bujard (1988) and Ptashne, et al.

#### **7.4.2.3.6 The Recombinase/Excision Sequence System**

The recombinase/excision sequence system can be any one that selectively removes DNA in a plant genome. The excision sequences are preferably unique in the plant, so that unintended cleavage of the plant genome does not occur. Several examples of such systems are discussed in Sauer, U.S. Pat. No. 4,959,317 and in Sadowski (1993). A preferred system is the bacteriophage CRE/LOX system, wherein the CRE protein performs site-specific recombination of DNA at LOX sites. Other systems include the resolvases (Hall, 1993), FLP (Pan, et al., 1993), SSV1 encoded integrase (Muskheishvili, et al., 1993), and the maize Ac/Ds transposon system (Shen and Hohn, 1992).

### **7.4.3 CONFERRING VIRAL RESISTANCE TO THE PLANT**

To practice the present invention, a viral gene must be isolated from the viral genome and inserted into a vector containing the genetic regulatory sequences necessary to express the inserted gene. Accordingly, a vector must be constructed to provide the regulatory sequences such that they will be functional upon inserting a desired gene. When the expression vector/insert construct is assembled, it is used to transform plant cells which are then used to regenerate plants. These transgenic plants carry the viral gene in the expression vector/insert construct. The gene is expressed in the plant and increased resistance to viral infection is conferred thereby.

#### **7.4.3.1 Isolation of a Viral Gene**

Several different courses exist to isolate a viral gene. To do so, one having ordinary skill in the art can use information about the genomic organization of potyviruses, cucumoviruses or comoviruses to locate and isolate the coat protein gene or the nuclear inclusion body genes. The coat protein gene in potyviruses, is located at the 3' end of the RNA, just prior to a stretch of about 200-300 adenine nucleotide residues. The nuclear inclusion body B (NIb) gene is located just 5' to the coat protein gene, and the nuclear inclusion body A (NIa) gene is 5' to the NIb gene. Additionally, the information related to proteolytic cleavage sites is used to determine the N- terminus of the potyvirus coat protein gene and the N- and C-terminus of non-coat protein genes. The protease recognition sites are conserved in the potyviruses and have been determined to be either the dipeptide Gln-Ser, Gln-Gly or Gln-Ala. The nucleotide sequences which encode these dipeptides can be determined.

#### **7.4.3.2 Insertion of the Viral Gene into a Vector**

Using methods well known in the art, a quantity of virus is grown and harvested. The viral RNA is then separated and the viral gene isolated using a number of known procedures. A cDNA library is created using the viral RNA, by methods known to the art. The viral RNA is incubated with primers that hybridize to the viral RNA and reverse transcriptase, and a complementary DNA molecule is produced. A DNA complement of the complementary DNA molecule is produced and that sequence represents a DNA copy (cDNA) of the original viral RNA molecule. The DNA complement can be produced in a manner that results in a single double stranded cDNA or polymerase chain reactions can

be used to amplify the DNA encoding the cDNA with the use of oligomer primers specific for the viral gene. These primers can include in addition to viral specific sequences, novel restriction sites used in subsequent cloning steps.

Thus, a double stranded DNA molecule is generated which contains the sequence information of the viral RNA. These DNA molecules can be cloned in *E. coli* plasmid vectors after the additions of restriction enzyme linker molecules by DNA ligase. The various fragments are inserted into cloning vectors, such as well-characterized plasmids, which are then used to transform *E. coli* to create a cDNA library.

Since potyvirus genes are generally conserved, oligonucleotides based on an analogous gene from a previous isolate or an analogous gene fragment from a previous isolate can be used as a hybridization probe to screen the cDNA library to determine if any of the transformed bacteria contain DNA fragments with the appropriate viral sequences. The cDNA inserts in any bacterial colonies which hybridize to these probes can be sequenced. The viral gene is present in its entirety in colonies which have sequences that extend 5' to sequences which encode a N-terminal proteolytic cleavage site and 3' to sequences which encode a C-terminal proteolytic cleavage site for the gene of interest.

Alternatively, cDNA fragments may be inserted in the sense orientation into expression vectors. Antibodies against a viral protein may be used to screen the cDNA expression library and the gene can be isolated from colonies which express the protein.

#### **7.4.3.3 Table of Selected Literature References to Methods of Isolating, cloning and Expressing Viral Genes**

The nucleotide sequences encoding the coat protein genes and nuclear inclusion genes of a number of viruses have been determined and the genes have been inserted into expression vectors. The expression vectors contain the necessary genetic regulatory sequences for expression of an inserted gene. The coat protein gene is inserted such that those regulatory sequences are functional and the genes can be expressed when incorporated into a plant genome. Selected literature references to methods of isolating, cloning and expressing viral genes are listed on Table 3, below.

**TABLE 3.**  
**Cloned Genes From RNA Viruses**

<b>Viral Gene</b>	<b>Reference</b>
Papaya ringspot cp	M.M. Fitch et al., Bio/Technology, 10, 1466(1992)
Potato virus X cp	K. Ling et al., Bio/Technology, 2, 752 (1991); A. Hoekema et al., Bio/Technology, 7, 273 (1989)
Watermelon Mosaic Virus II cp	H. Quemada et al., J. Gen. Virol., 71, 1451 (1990); S. Namba et al., Phytopathology, 82, 940 (1992)
Zucchini yellow MosaicVirus cp	S. Namba et al., Phytopathology, 82, 940 (1992)
Tobacco Mosaic Virus cp	R.S. Nelson et al., Bio/Technology, 6, 403 (1988); P. Powell Abel et al., Science, 232, 738 (1986)
Alfalfa Mosaic Virus cp	Loesch-Fries et al., EMBO J., 6, 1845 (1987); N.E. Turner et al., EMBO J., 6, 1181 (1987)
Soybean Mosaic Virus cp	D.M. Stark et al., Biotechnology, 7, 1257 (1989)
Cucumber Mosaic Virus strain C cp	H.Q. Quemada et al., Molec. Plant Pathol., 81, 794 (1991)
Cucumber Mosaic Virus strain WL cp	UpJohn Co. (PCT W090/02185)
Tobacco etch virus cp	Allison et al., Virology, 147, 309 (1985)
Tobacco etch virus nuclear inclusion protein	J.C. Carrington et al., J. Virol., 61, 2540 (1987)
Pepper Mottle Virus cp	W.G. Dougherty et al., Virology, 146, 282 (1985)
Potato virus Y cp	D.D. Shukla et al., Virology, 152, 118 (1986)

Potato virus Y nuclear inclusion protein	European Patent Application 578,627
Potato virus X cp	C. Lawson et al., Biotechnology, 8, 127 (1990)
Tobacco streak virus (TSV) cp	C.M. Van Dun et al., Virology, 164, 383 (1988)

#### **7.4.4 FORMATION OF A DNA CONSTRUCT**

##### **7.4.4.1 Fusion Gene Including a Trait DNA Molecule and a Silencer DNA Molecule**

The present invention is directed to a DNA construct formed from a fusion gene which includes a trait DNA molecule and a silencer DNA molecule. The trait DNA molecule has a length that is insufficient to impart a desired trait to plants transformed with the trait DNA molecule. The silencer DNA molecule is operatively coupled to the trait DNA molecule with the trait and silencer DNA molecules collectively having sufficient length to impart the trait to plants transformed with the DNA construct.

##### **7.4.4.2 Plurality of Trait DNA Molecules**

In an alternative embodiment of the present invention, the DNA construct can be a fusion gene comprising plurality of trait DNA molecules at least some of which having a length that is insufficient to impart that trait to plants transformed with that trait DNA molecule. However, the plurality of trait DNA molecules collectively have a length sufficient to impart the traits to plants transformed with the DNA construct and to effect post-transcriptional silencing of the fusion gene.

A particularly preferred aspect of the present invention is where the DNA construct includes a plurality of trait DNA molecules each having a length insufficient to impart the trait to plants transformed with the trait DNA molecule alone. It is also possible for some of the trait DNA molecules to have a length sufficient to impart their respective trait; however, not all such DNA molecules will have such a length.

##### **7.4.4.3 Transformed Plant Cells With Expression Constructs For Production Of Mature Proteins**

Plant cells or tissues derived from the members of the family known as the Gramineae are transformed with expression constructs (i.e., plasmid DNA into which the gene of interest has been inserted) using a variety of standard techniques (e.g., electroporation, protoplast fusion or microparticle bombardment). The expression construct includes a transcription regulatory region (promoter) whose transcription is specifically upregulated by the presence or absence of a small molecule, such as the reduction or depletion of sugar, e.g., sucrose, in culture medium, or in plant tissues, e.g.,



germinating seeds. In the present invention, particle bombardment is the preferred transformation procedure.

The construct also includes a gene encoding a mature heterologous protein in a form suitable for secretion from plant cells. The gene encoding the recombinant heterologous protein is placed under the control of a metabolically regulated promoter. Metabolically regulated promoters are those in which mRNA synthesis or transcription, is repressed or upregulated by a small metabolite or hormone molecule, such as the rice RAmy3D and RAmy3E promoters, which are upregulated by sugar-depletion in cell culture. For protein production in germinating seeds from regenerated transgenic plants, a preferred promoter is the-Ramy 1A promoter, which is up-regulated by gibberellic acid during seed germination. The expression construct also utilizes additional regulatory DNA sequences e.g., preferred codons, termination sequences, to promote efficient translation of AAT, as will be described.

#### 7.4.5 DISEASE RESISTANCE

One aspect of the present invention relates to the use of trait DNA molecules which are heterologous to the plant -- e.g., DNA molecules that confer disease resistance to plants transformed with the DNA construct. The present invention is useful in plants for imparting resistance to a wide variety of pathogens including viruses, bacteria, fungi, viroids, phytoplasmas, nematodes, and insects. The present invention may also be used in mammals to impart genetic traits. Resistance, inter alia, to the following viruses can be achieved by the method of the present invention: tomato spotted wilt virus, impatiens necrotic spot virus, groundnut ringspot virus, potato virus Y, potato virus X, tobacco mosaic virus, turnip mosaic virus, tobacco etch virus, papaya ringspot virus, tomato mottle virus, tomato yellow leaf curl virus, or-combinations thereof. Resistance, inter alia, to the following bacteria can also be imparted to plants in accordance with present invention: *Pseudomonas solanacearum*, *Pseudomonas syringae* pv. *tabaci*, *Xanthomonas campestris* pv. *pelargonii*, and *Agrobacterium tumefaciens*. Plants can be made resistant, inter alia, to the following fungi by use of the method of the present invention: *Fusarium oxysporum* and *Phytophthora infestans*. Suitable DNA molecules include a DNA molecule encoding a coat protein, a replicase, a DNA molecule not encoding protein, a DNA molecule encoding a viral gene product, or combinations thereof.

##### 7.4.5.1 Sense/Antisense orientation

The DNA molecule conferring disease resistance can be positioned within the DNA construct in sense orientation. Alternatively, it can have an antisense orientation. Antisense RNA technology involves the production of an RNA molecule that is complementary to the messenger RNA molecule of a target gene; the antisense RNA can potentially block all expression of the targeted gene. In the anti-virus context, plants are made to express an antisense RNA molecule corresponding to a viral RNA (that is, the antisense RNA is an RNA molecule which is complementary to a plus sense RNA species encoded by an infecting virus). Such plants may show a slightly decreased susceptibility to infection by that virus. Such a complementary RNA molecule is termed antisense RNA.

#### **7.4.6 OTHER TRAITS**

The present invention is also used to confer traits other than disease resistance on plants. For example, DNA molecules which impart a plant genetic trait can be used as the DNA trait molecule of the present invention. In this aspect of the present invention, suitable trait DNA molecules encode for desired color, enzyme production, or combinations thereof.

##### **7.4.6.1 GENE SILENCING**

The silencer DNA molecule of the present invention can be selected from virtually any nucleic acid which effects gene silencing. This involves the cellular mechanism to degrade mRNA homologous to the transgene mRNA. The silencer DNA molecule can be heterologous to the plant, need not interact with the trait DNA molecule in the plant, and can be positioned 3' to the trait DNA molecule. For example, the silencer DNA molecule can be a viral cDNA molecule, a jellyfish green fluorescence protein encoding DNA molecule, a plant DNA molecule, or combinations thereof.

While not wishing to be bound by theory, by use of the construct of the present invention, it is believed that post-transcriptional gene silencing is achieved. More particularly, the silencer DNA molecule is believed to boost the level of heterologous RNA within the cell above a threshold level. This activates the degradation mechanism by which viral resistance is achieved.

##### **7.4.6.1.1 TRAIT & SILENCER DNA MOLECULES ENCODING RNA MOLECULES - TRANSLATABLE**

It is possible for the DNA construct of the present invention to be configured so that the trait and silencer DNA molecules encode RNA molecules which are translatable. As a result, that RNA molecule will be translated at the ribosomes to produce the protein encoded by the DNA construct. Production of proteins in this manner can be increased by joining the cloned gene encoding the DNA construct of interest with synthetic double-stranded oligonucleotides which represent a viral regulatory sequence (i.e., a 5' untranslated sequence). See U.S. Patent No. 4,820,639 to Gehrke and U.S. Patent No. 5,849,527 to Wilson which are hereby incorporated by reference.

#### **7.4.6.1.2 TRAIT & SILENCER DNA MOLECULES ENCODING RNA MOLECULES – NOT TRANSLATABLE**

Alternatively, the DNA construct of the present invention can be configured so that the trait and silencer DNA molecules encode mRNA which is not translatable. This is achieved by introducing into the DNA molecule one or more premature stop codons, adding one or more bases (except multiples of 3 bases) to displace the reading frame, removing the translation initiation codon, etc. See U.S. Patent No. 5,583,021 to Dougherty et al., which is hereby incorporated by reference.

#### **7.4.7 RECOMBINANT DNA TECHNOLOGY**

The subject DNA construct can be incorporated in cells using conventional recombinant DNA technology. Generally, this involves inserting the DNA construct into an expression system to which the DNA construct is heterologous (i.e. not normally present). The heterologous DNA construct may be inserted into the expression system or vector in proper sense orientation and correct reading frame. The vector contains the necessary elements for the transcription of the inserted sequences.

##### **7.4.7.1 Restriction enzyme cleavage and ligation**

U.S. Patent No. 4,237,224 to Cohen and Boyer, which is hereby incorporated by reference, describes the production of expression systems in the form of recombinant plasmids using restriction enzyme cleavage and ligation with DNA ligase. These recombinant plasmids are then introduced by means of transformation and replicated in unicellular cultures including procaryotic organisms and eucaryotic cells grown in tissue culture.

##### **7.4.7.2 Introduction to viruses**

Recombinant genes may also be introduced into viruses, such as vaccinia virus. Recombinant viruses can be generated by transfection of plasmids into cells infected with virus.

#### **7.4.8 INCORPORATION INTO A HOST CELL**

Once the DNA construct has been cloned into an expression system, it is ready to be incorporated into a host cell. Such incorporation can be carried out by the various forms of transformation noted above, depending upon the vector/host cell system. Suitable host cells include, but are not limited to, bacteria, virus, plant, is and the like cells.

##### **7.4.8.1 PLANT TRANSFORMATION - PRODUCTION OF MATURE PROTEINS IN PLANTS**

Expression vectors for use in the present invention comprise a chimeric gene (or expression cassette), designed for operation in plants, with companion sequences upstream and downstream from the expression cassette. The companion sequences will be of plasmid or viral origin and provide necessary characteristics to the vector to permit the vectors to move DNA from bacteria to the desired plant host. For transformation of plants, the chimeric gene is placed in a suitable expression vector designed for operation in plants. The vector includes suitable elements of plasmid or viral origin that provide necessary characteristics to the vector to permit the vectors to move DNA from bacteria to the desired plant host. Suitable components of the expression vector, including an inducible promoter, coding sequence for a signal peptide, coding sequence for a mature heterologous protein, and suitable termination sequences, are discussed below. One exemplary vector is the p3Dvl.O (p3D(AAT)v1.0) vector described herein.

###### **7.4.8.1.1 Transformation Vector**

Vectors containing a chimeric gene of the present invention may also include selectable markers for use in plant cells (such as the nptII kanamycin resistance gene, for selection in kanamycin-containing or the phosphinothricin acetyltransferase gene, for selection in medium containing phosphinothricin (PPT).

The vectors may also include sequences that allow their selection and propagation in a secondary host, such as sequences containing an origin of replication and a selectable marker such as antibiotic or herbicide resistance genes, e.g., HPH (Hagio et al., Plant Cell Reports 14:329 (1995); van der Elzer, Plant Mol. Biol. 5:299-302 (1985)). Typical secondary hosts include bacteria and yeast. In one embodiment, the secondary host is *Escherichia coli*, the origin of replication is a *colEI*-type, and the selectable marker is a

gene encoding ampicillin resistance. Such sequences are well known in the art and are commercially available as well (e.g., Clontech, Palo Alto, CA; Stratagene, La Jolla, CA).

The vectors of the present invention may also be modified to intermediate plant transformation plasmids that contain a region of homology to an Agrobacterium tumefaciens vector, a T-DNA border region from Agrobacterium tumefaciens, and chimeric genes or expression cassettes (described above). Further, the vectors of the invention may comprise a disarmed plant tumor inducing plasmid of Agrobacterium tumefaciens.

## **7.4.9 PLANT EXPRESSION VECTOR PRODUCTION OF MATURE PROTEINS IN PLANTS**

### **7.4.9.1 SUITABLE VECTORS**

Suitable vectors include, but are not limited to, the following viral vectors such as lambda vector system gt11, gt WES.tB, Charon 4, and plasmid vectors such as pER322, pBR325, pACYC177, pACYC1084, pUC8, pUC9, pUC18, pUC19, pLG339, pR290, pKC37, pKC101, SV 40, pBluescript II SK +/- or KS +/- (see "Stratagene Cloning Systems" Catalog (1993) from Stratagene, La Jolla, Calif, which is hereby incorporated by reference), pQE, pIH821, pGEX, pET series (see F.W. Studier et. al., "Use of T7 RNA Polymerase to Direct Expression of Cloned Genes," Gene Expression Technolog vol. 185 (1990), which is hereby incorporated by reference), and any derivatives thereof. Recombinant molecules can be introduced into cells via transformation, particularly transduction, conjugation, mobilization, or electroporation. The DNA sequences are cloned into the vector using standard cloning procedures in the art, as described by Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Springs Laboratory, Cold Springs Harbor, New York (1989), which is hereby incorporated by reference.

### **7.4.9.2 HOST-VECTOR SYSTEMS**

A variety of host-vector systems may be utilized to carry out the present invention. Primarily, the vector system must be compatible with the host cell used. Host-vector systems include but are not limited to the following: bacteria transformed with bacteriophage DNA, plasmid DNA, or cosmid DNA; microorganisms such as yeast containing yeast vectors; mammalian cell systems infected with virus (e.g., vaccinia virus, adenovirus, etc.); insect cell systems infected with virus (e.g., baculovirus); and plant cells infected by bacteria. The expression elements of these vectors vary in their strength and specificities. Depending upon the host- vector system utilized, any one of a number of suitable transcription and, perhaps, translation elements can be used.



#### **7.4.10 GENETIC SIGNALS AND PROCESSING EVENTS**

In order to express the viral gene, the necessary genetic regulatory sequences must be provided. Since the proteins encoded in a potyvirus, genome are produced by the post translational processing of a polyprotein, a viral gene isolated from viral RNA does not contain transcription and translation signals necessary for its expression once transferred and integrated into a plant genome. It must, therefore, be engineered to contain a plant expressible promoter, a translation initiation codon (ATG) and a plant functional poly(A) addition signal (AATAAA) 3' of its translation termination codon. In the present invention, a viral gene is inserted into a vector which contains cloning sites for insertion 3' of the initiation codon and 5' of the poly(A) signal. The promoter is 5' of the initiation codon such that when structural genes are inserted at the cloning site, a functional unit is formed in which the inserted genes are expressed under the control of the various genetic regulatory sequences.

Different genetic signals and processing events control many levels of gene expression (e.g., DNA transcription and messenger RNA (mRNA) translation).

##### **7.4.10.1 Signal Sequences for production of mature proteins**

In addition to encoding the protein of interest, the chimeric gene encodes a signal sequence (or signal peptide) that allows processing and translocation of the protein, as appropriate. Suitable signal sequences are described in above-referenced PCT application WO 95/14099. The plant signal sequence is placed in frame with a heterologous nucleic acid encoding a mature protein, forming a construct which encodes a fusion protein having an N- terminal region corresponding to the signal peptide and, immediately adjacent to the C- terminal amino acid of the signal peptide, the N-terminal amino acid of the mature heterologous protein. The expressed fusion protein is subsequently secreted and processed by signal peptidase cleavage precisely at the junction of the signal peptide and the mature protein, to yield the mature heterologous protein.

In another embodiment of the invention, the coding sequence in the fusion protein gene, in at least the coding region for the signal sequence, may be codon-optimized for optimal expression in plant cells, e.g., rice cells, as described below.

#### 7.4.10.2 PROMOTORS

##### Transcription dependent upon the presence of a promotor

Transcription of DNA is dependent upon the presence of a promotor which is a DNA sequence that directs the binding of RNA polymerase and thereby promotes mRNA synthesis. The DNA sequences of eucaryotic promotors differ from those of procaryotic promotors. Furthermore, eucaryotic promotors and accompanying genetic signals may not be recognized in or may not function in a procaryotic system, and, further, procaryotic promotors are not recognized and do not function in eucaryotic cells.

The segment of DNA referred to as the promoter is responsible for the regulation of the transcription of DNA into mRNA. A number of promoters which function in plant cells are known in the art and may be employed in the practice of the present invention. These promoters may be obtained from a variety of sources such as plants or plant viruses, and may include but are not limited to promoters isolated from the caulimovirus group such as the cauliflower mosaic virus 35S promoter (CaMV35S), the enhanced cauliflower mosaic virus 35S promoter (enh CaMV35S), the figwort mosaic virus full-length transcript promoter (FMV35S), and the promoter isolated from the chlorophyll alb binding protein. Other useful promoters include promoters which are capable of expressing the potyvirus proteins in an inducible manner or in a tissue-specific manner in certain cell types in which the infection is known to occur. For example, the inducible promoters from phenylalanine ammonia lyase, chalcone synthase, hydroxyproline rich glycoprotein, extensin, pathogenesis-related proteins (e.g. PR-1a), and wound-inducible protease inhibitor from potato may be useful.

Preferred promoters for use in the present viral gene expression cassettes include the constitutive promoters from CaMV, the Ti genes nopaline synthase (Bevan et al., Nucleic Acids Res. II, 369-385 (1983)) and octopine synthase (Depicker et al., J. Mol. Appl. Genet., 1, 561- 564 (1982)), and the bean storage protein gene phaseolin. The poly(A) addition signals from these genes are also suitable for use in the present cassettes. The particular promoter selected is preferably capable of causing sufficient expression of the DNA coding sequences to which it is operably linked, to result in the production of amounts of the proteins or the RNAs effective to provide viral resistance, but not so much as to be detrimental to the cell in which they are expressed. The promoters selected should

be capable of functioning in tissues including but not limited to epidermal, vascular, and mesophyll tissues. The actual choice of the promoter is not critical, as long as it has sufficient transcriptional activity to accomplish the expression of the preselected proteins or antisense RNA, and subsequent conferral of viral resistance to the plants.

Promoters vary in their "strength" (i.e. their ability to promote transcription). For the purposes of expressing a cloned gene, it is desirable to use strong promoters in order to obtain a high level of transcription and, hence, expression of the gene. Depending upon the host cell system utilized, any one of a number of suitable promoters may be used. For instance, when cloning in *E. coli*, its bacteriophages, or plasmids, promoters such as the T7 phage promoter, lac promoter, trp promoter, recA promoter, ribosomal RNA promoter, the  $P_R$  and  $P_L$  promoters of coliphage lambda and others, including but not limited, to lacUV5, ompF, bla, lpp, and the like, may be used to direct high levels of transcription of adjacent DNA segments. Additionally, a hybrid trp-lacUV5 (tac) promoter or other *E. coli* promoters produced by recombinant DNA or other synthetic DNA techniques may be used to provide for transcription of the inserted gene.

#### **7.4.10.2.1 Promoters that transcribe the cereal $\alpha$ -amylase genes and sucrose synthase genes**

The transcription regulatory or promoter region is chosen to be regulated in a manner allowing for induction under selected cultivation conditions, e.g., sugar depletion in culture or water uptake followed by gibberellic acid production in germinating seeds. Suitable promoters, and their method of selection are detailed in above-cited PCT application WO 95/14099. Examples of such promoters include those that transcribe the cereal  $\alpha$ -amylase genes and sucrose synthase genes, and are repressed or induced by small molecules, like sugars, sugar depletion or phytohormones such as gibberellic acid or abscisic acid. Representative promoters include the promoters from the rice  $\alpha$ -amylase RAm1A, RAm1B, RAm2A, RAm3A, RAm3B, RAm3C, RAm3D, and RAm3E genes, and from the pM/C, gKAmy14I, gKAmy155, Amy32b, and HV18 barley ( $\alpha$ -amylase genes. These promoters are described, for example, in ADVANCES IN PLANT BIOTECHNOLOGY Ryu, D.D.Y., et al, Eds., Elsevier, Amsterdam, 1994, p.37, and references cited therein. Other suitable promoters include the sucrose synthase and sucrose-6-phosphate-synthetase (SPS) promoters from rice and barley.

Other suitable promoters include promoters which are regulated in a manner allowing for induction under seed-maturation conditions. Examples of such promoters include those associated with the following monocot storage proteins: rice glutelins, oryzins, and prolamines, barley hordeins, wheat gliadins and glutelins, maize zeins and glutelins, oat glutelins, and sorghum kafirins, millet pennisetins, and rye secalins.

A preferred promoter for expression in germinating seeds is the rice  $\alpha$ -amylase RAmy1A promoter, which is upregulated by gibberellic acid. Preferred promoters for expression in cell culture are the rice  $\alpha$ -amylase RAmy3D and RAmy3E promoters which are strongly upregulated by sugar depletion in the culture. These promoters are also active during seed germination. A preferred promoter for expression in maturing seeds is the barley endosperm-specific BI-hordein promoter (Brandt, A., et al., (1985) Carlsberg Res. Commun. 50:333-345).

The chimeric gene may further include, between the promoter and coding sequences, the 5' untranslated region (5' UTR) of an inducible monocot gene, such as the 5' UTR derived from one of the rice or barley  $\alpha$ -amylase genes mentioned above. One preferred 5' UTR is that derived from the RAmy1A gene, which is effective to enhance the stability of the gene transcript.

#### **7.4.10.2.2 Use of inducers**

Bacterial host cell strains and expression vectors may be chosen which inhibit the action of the promoter unless specifically induced. In certain operations, the addition of specific inducers is necessary for efficient transcription of the inserted DNA. For example, the lac operon is induced by the addition of lactose or IPTG (isopropylthio-beta-D-galactoside). A variety of other operons, such as trp, pro, etc., are under different controls.

#### **7.4.10.3 TRANSCRIPTION INITIATION SIGNALS**

Specific initiation signals are also required for efficient gene transcription in procaryotic cells.

These transcription initiation signals may vary in "strength" as measured by the quantity of gene specific messenger RNA and protein synthesized, respectively. The DNA expression vector, which contains a promotor, may also contain any combination of various "strong" transcription initiation signals.

#### **7.4.10.4 Translation dependent on Shine-Dalgarno sequence (in procaryotes)**

Similarly, translation of mRNA in procaryotes depends upon the presence of the proper procaryotic signals which differ from those of eucaryotes. Efficient translation of mRNA in procaryotes requires a ribosome binding site called the Shine-Dalgarno ("SD") sequence on the mRNA. This sequence is a short nucleotide sequence of mRNA that is located before the start codon, usually AUG, which encodes the amino-terminal methionine of the protein. The SD sequences are complementary to the 3'-end of the 16S rRNA (ribosomal RNA) and probably promote binding of mRNA to ribosomes by duplexing with the rRNA to allow correct positioning of the ribosome. For a review on maximizing gene expression, see Roberts and Lauer, *Methods in Enzymology*, 68:473 (1979), which is hereby incorporated by reference.

The non-translated leader sequence can be derived from any suitable source and can be specifically modified to increase the translation of the mRNA. The 5' non-translated region can be obtained from the promoter selected to express the gene, an unrelated promoter, the native leader sequence of the gene or coding region to be expressed, viral RNAs, suitable eucaryotic genes, or a synthetic gene sequence. The present invention is not limited to the constructs presented in the following examples.

#### **7.4.10.5 Naturally-Occurring, Heterologous Protein Coding Sequences**

**7.4.10.5.1 (i)  $\alpha_1$ -Antitrypsin:** Mature human AAT is composed of 394 amino acids. The protein has N-glycosylation sites at asparagines 46, 83 and 247.

**7.4.10.5.2 (ii) Antithrombin III:** Mature human ATIII is composed of 432 amino acids. The protein has N-glycosylation sites at the four asparagine residues 96, 135, 155, and 192.

**7.4.10.5.3 (iii) Human serum albumin:** Mature HSA as found in human serum is composed of 585 amino acids. The protein has no N-linked glycosylation sites.

**7.4.10.5.4 (iv) Subtilisin BPN':** Native proBPN' as produced in *B. ainyloliquefaciens* is composed of 352 amino acids. The proBPN' polypeptide contains a 77 amino acid "pro" moiety. The remainder of the polypeptide, which forms the mature active BPN', is a 275 amino acid sequence. Native l3PN' as produced in *Bacillus* is not glycosylated.

#### **7.4.10.6 A4. Codon-Optimized Coding Sequences**

In accordance with one aspect of the invention, it has been discovered that a severalfold enhancement of expression level can be achieved in plant cell culture by modifying the native coding sequence of a heterologous gene by contain predominantly or exclusively, highest-frequency codons found in the plant cell host.

The method will be illustrated for expression of a heterologous gene in rice plant cells, it being recognized that the method is generally applicable to any monocot. As a first step, a representative set of known coding gene sequence from rice is assembled. The sequences are then analyzed for codon frequency for each amino acid, and the most frequent codon is selected for each amino acid. This approach differs from earlier reported codon matching methods, in which more than one frequent codon is selected for at least some of the amino acids. The optimal codons selected in this manner for rice and barley are shown in Table 4.

**TABLE 4**

<b>Amino Acid</b>	<b>Rice Preferred Codon</b>	<b>Barley Preferred Codon</b>
Ala A	GCC	
Arg R	CGC	

Asn N	AAC	
Asp D	GAC	
Cys C	UGC	
Gln Q	CAG	
Glu E	GAG	
Gly G	GGC	
His H	CAC	
Ile I	AUC	
Leu L	CUC	
Lys K	AAG	
Phe F	UUC	
Pro P	CCG	CCC
Ser S	AGC	UCC
Thr T	ACC	
Tyr Y	UAC	
Val V	GUC	GUG
stop	UAA	UGA

As indicated above, the fusion protein coding sequence in the chimeric gene is constructed such that the final (C-terminal) codon in the signal sequence is immediately followed by the codon for the N-terminal amino acid in the mature form of the heterologous protein.

In a preferred embodiment, the BPN' coding sequence is further modified to eliminate potential N-glycosylation sites, as native BPN' is not glycosylated. Table 5 illustrates preferred codon substitutions, which eliminate all potential N-glycosylation sites in subtilisin BPN'.

**TABLE 5**

N-Glycosylation Sites	Location (Asn) (in mature protein)	Amino Acid Substitution
-----------------------	------------------------------------	-------------------------

Asn Asn Ser	61	Thr Asn Ser
Asn Asn Ser	76	Thr Asn Ser
Asn Met Ser	123	Thr Met Ser
Asn Gly Thr	218	Ser Gly Thr'
Asn Trp Thr	240	Thr Trp Thr

'improved thermostability; Bryan, et al., Proteins: Structure, Function, and Genetics 1:326 (1986).

#### 7.4.10.7 TERMINATION SEQUENCE COUPLED TO THE FUSION END

The present invention can also utilize a termination sequence operatively coupled to the fusion gene to end transcription. Suitable transcription termination sequences include the termination region of a 3' non-translated region. This will cause the termination of transcription and the addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequence. The termination region or 31 non-translated region will be additionally one of convenience. The termination region may be native with the promoter region or may be derived from another source, and preferably includes a terminator and a sequence coding for polyadenylation. Suitable 3' non-translated regions include but are not limited to: (1) the 3' transcribed, non-translated regions containing the polyadenylated signal of *Agrobacterium* tumor-inducing (Ti) plasmid genes, such as the nopaline synthase (NOS) gene or the 35S promoter terminator gene, and (2) plant genes like the soybean 7S storage protein genes and the pea small subunit of the ribulose 1,5-bisphosphate carboxylase-oxygenase (ssRUBISCO) E9 gene.

The termination region or 3' non-translated region which is employed is one which will cause the termination of transcription and the addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequence. The termination region may be native with the promoter region, native with the structural gene, or may be derived from another source, and preferably include a terminator and a sequence coding for polyadenylation. Suitable 3' non-translated regions of the chimeric plant gene include but are not limited to: (1) the 3' transcribed, non-translated regions containing the polyadenylation signal of *Agrobacterium* tumor-inducing (Ti) plasmid genes, such as the nopaline synthase (NOS) gene, and (2) plant genes like the soybean 7S storage protein genes.



#### 7.4.10.8 Transcription and Translation Terminators for production of mature proteins

The chimeric gene may also include, downstream of the coding sequence, the 3' untranslated region (3' UTR) from an inducible monocot gene, such as one of the rice or barley  $\alpha$ -amylase genes mentioned above. One preferred 3' UTR is that derived from the RAmYL gene. This sequence includes non-coding sequence 5' to the polyadenylation site, the polyadenylation site, and the transcription termination sequence. The transcriptional termination region may be selected, particularly for stability of the mRNA to enhance expression. Polyadenylation tails (Alber and Kawasaki, 1982, Mol. and Appl. Genet. 1:419-434) are also commonly added to the expression cassette to optimize high levels of transcription and proper transcription termination, respectively. Polyadenylation sequences include but are not limited to the *Agrobacterium* octopine synthetase signal (Gielen, et al., EMBO J. 3:835- 846 (1984) or the nopaline synthase of the same species (Depicker, et al. , Mol. Appl. Genet. 1:561- 573 (1982)).

Since the ultimate expression of the heterologous protein will be in a eukaryotic cell (in this case, a member of the grass family), it is desirable to determine whether any portion of the cloned gene contains sequences which will be processed out as introns by the host's splicing machinery. If so, site-directed mutagenesis of the "intron" region may be conducted to prevent losing a portion of the genetic message as a false intron code (Reed and Maniatis, Cell 41:95-105 (1985)).

#### 7.4.11 SELECTABLE MARKER GENE

Selectable marker genes may be incorporated into the present expression cassettes and used to select for those cells or plants which have become transformed. The marker gene employed may express resistance to an antibiotic, such as kanamycin, gentamycin, G418, hygromycin, streptomycin, spectinomycin, tetracycline, chloramphenicol, and the like.

Other markers could be employed in addition to or in the alternative, such as, for example, a gene coding for herbicide tolerance such as tolerance to glyphosate, sulfonylurea, phosphinothricin, or bromoxynil. Additional means of selection could include resistance to methotrexate, heavy metals, complementation providing prototrophy to an auxotrophic host, and the like.

For example, see Table 1 of PCT WO/91/10725, cited above. The present invention also envisions replacing all of the virus-associated genes with an array of selectable marker genes.

The particular marker employed will be one which will allow for the selection of transformed cells as opposed to those cells which were not transformed. Depending on the number of different host species one or more markers may be employed, where different conditions of selection would be useful to select the different host, and would be known to those of skill in the art. A screenable marker or "reporter gene" such as the 0-glucuronidase gene or luciferase gene may be used in place of, or with, a selectable marker. Cells transformed with this gene may be identified by the production of a blue product on treatment with 5-bromo-4-chloro-3-indoyl- $\beta$ -D-glucuronide (X-Gluc).

In developing the present expression construct, the various components of the expression construct such as the DNA sequences, linkers, or fragments thereof will normally be insetted into a convenient cloning vector, such as a plasmid or phage, which is capable of replication in a bacterial host, such as *E. coli*. Numerous cloning vectors exist that have been described in the literature. After each cloning, the cloning vector may be isolated and subjected to further manipulation, such as restriction, insertion of new fragments, ligation, deletion, resection, insertion, in vitro mutagenesis, addition of

polylinker fragments, and the like, in order to provide a vector which will meet a particular need.

## **7.4.12 TRANSFERRING RECOMBINANT DNA INTO PLANT CELL**

### **7.4.12.1 Use of micropipettes or polyethylene glycol**

In producing transgenic plants, the DNA construct in a vector described above can be microinjected directly into plant cells by use of micropipettes to transfer mechanically the recombinant DNA. Crossway, *Mol. Gen. Genetics*, 202:179-85 (1985), which is hereby incorporated by reference. The genetic material may also be transferred into the plant cell using polyethylene glycol. Krens, et al., *Nature*, 296:72-74 (1982), which is hereby incorporated by reference.

### **7.4.12.2 Particle Bombardment (Biolistic Transformation)**

Another approach to transforming plant cells with the DNA construct is particle bombardment (also known as biolistic transformation) of the host cell. This can be accomplished in one of several ways. The first involves propelling inert or biologically active particles at cells. This technique is disclosed in U.S. Patent Nos. 4,945,050, 5,036,006, and 5,100,792, all to Sanford et al., which are hereby incorporated by reference. Generally, this procedure involves propelling inert or biologically active particles at the cells under conditions effective to penetrate the outer surface of the cell and to be incorporated within the interior thereof. When inert particles are utilized, a vector containing the DNA construct can be introduced into the cell by coating the particles with the vector containing that heterologous DNA construct. Alternatively, the target cell can be surrounded by the vector so that the vector is carried into the cell by the wake of the particle. Biologically active particles (e.g., dried bacterial cells containing the vector and heterologous DNA construct) can also be propelled into plant cells.

### **7.4.12.2 Fusion of protoplasts with other entities**

Yet another method of introduction is fusion of protoplasts with other entities, - either minicells, cells, lysosomes or other fusible lipid-surfaced bodies. Fraley, et al., *Proc. Natl. Acad. Sci. USA*, 79:1859-63 (1982), which is hereby incorporated by reference.

### **7.4.12.4 Electroporation**

The DNA molecule may also be introduced into the plant cells by electroporation. Fromm et al., *Proc. Natl. Acad. Sci. USA*, 82:5824 (1985), which is hereby incorporated by reference. In this technique, plant protoplasts are electroporated in the presence of

plasmids containing the expression cassette. Electrical impulses of high field strength reversibly permeabilize biomembranes allowing the introduction of the plasmids. Electroporated plant protoplasts reform the cell wall, divide, and regenerate.

#### 7.4.12.5 Infection with *Agrobacterium tumefaciens* or *A. rhizogenes*

Another method of introducing the DNA molecule into plant cells is to infect a plant cell with *Agrobacterium tumefaciens* or *A. rhizogenes* previously transformed with the gene. Under appropriate conditions known in the art, the transformed plant cells are grown to form shoots or roots, and develop further into plants. Generally, this procedure involves inoculating the plant tissue with a suspension of bacteria and incubating the tissue for 48 to 72 hours on regeneration medium without antibiotics at 25-28°C.

*Agrobacterium* is a representative genus of the gram-negative family Rhizobiaceae. Its species are responsible for crown gall (*A. tumefaciens*) and hairy root disease (*A. rhizogenes*). The plant cells in crown gall tumors and hairy roots are induced to produce amino acid derivatives known as opines, which are catabolized only by the bacteria. The bacterial genes responsible for expression of opines are a convenient source of control elements for chimeric expression cassettes. In addition, assaying for the presence of opines can be used to identify transformed tissue.

Heterologous genetic sequences can be introduced into appropriate plant cells, by means of the Ti plasmid of *A. tumefaciens* or the Ri plasmid of *A. rhizogenes*. The Ti or Ri plasmid is transmitted to plant cells on infection by *Agrobacterium* and is stably integrated into the plant genome. J. Schell, Science, 237:1176-83 (1987), which is hereby incorporated by reference.

For *Agrobacterium*-mediated transformation, the expression cassette will be included in a vector, and flanked by fragments of the *Agrobacterium* Ti or Ri plasmid, representing the right and, optionally the left, borders of the Ti or Ri plasmid transferred DNA (T-DNA). This facilitates integration of the present chimeric DNA sequences into the genome of the host plant cell. This vector will also contain sequences that facilitate replication of the plasmid in *Agrobacterium* cells, as well as in *E. coli* cells.

All DNA manipulations are typically carried out in *E. coli* cells, and the final plasmid bearing the potyvirus expression cassette is moved into Agrobacterium cells by direct DNA transformation, conjugation, and the like. These Agrobacterium cells will contain a second plasmid, also derived from Ti or Ri plasmids. This second plasmid will carry all the vir genes required for transfer of the foreign DNA into plant cells.

Suitable plant transformation cloning vectors include those derived from a Ti plasmid of Agrobacterium tumefaciens, as generally disclosed in Glassman et al. (U.S. Pat. No. 5,258,300). In addition to those disclosed, for example, Herrera-Estrella, *Nature*, 303, 209 (1983), *Biotechnica* (published PCT application PCT WO/91/10725), and U.S. patent 4,940,838, issued to Schilperoort et al.

#### 7.4.13 METHOD FOR MAKING GENETICALLY RECOMBINANT PLANTS IN COMMERCIALY FEASIBLE NUMBERS

In one preferred embodiment, the invention provides for a process for propagating plants by tissue culture in such a way as both to conserve desired plant morphology and to transform the plant with respect to one or more desired genes. The method includes the steps of (a) creating an Agrobacterium vector containing the gene sequence desired to be transferred to the propagated plant, preferably together with a marker gene; (b) taking one or more petiole explants from a mother plant and inoculating them with the Agrobacterium vector; (c) conducting callus formation in the petiole sections in culture, in the dark; and (d) culturing the resulting callus in growth medium having a benzylamino growth regulator such as benzylaminopurine or, most preferably, benzylaminopurineriboside. Additional optional growth regulators including auxins and cytokinins (indole butyric acid, benzylamine, benzyladenine, benzylaminopurine, alpha naphthylacetic acid and others known in the art) may also be present. Preferably, the petiole tissue is taken from *Pelargonium x domesticum* and the Agrobacterium vector contains an antisense gene for ACC synthase or ACC oxidase to prevent ACC synthase or ACC oxidase expression and, in turn, preventing ethylene formation. *Pelargoniums* propagated in culture using the present technique are resistant to wilting and petal shatter, and are morphologically conserved due to the use of petiole explants specifically and the particular culture media disclosed.

##### 7.4.13.1 Using Petioles as the Explant Tissue

Although in theory any anatomic explants can be mixed with Agrobacterium containing the desired gene sequences to be transferred, followed by tissue culture propagation of transgenic transformed plants, in practice we have encountered unexpectedly good results using petioles as the explant tissue. We have found that morphologic conservation is virtually assured with the use of leaf petiole tissue, whereas morphologic variation—even between two generations—can result when explants of other tissue, i.e. leaf tissue, are used. Moreover, the petiole explants should be taken from stock plants (mother plants) of which commercial propagation is desired. Commercial viability is attributable to the large number of transgenically transformed plants which can be produced from a relatively few petioles taken from the mother plant—particularly

because leaf petioles can be harvested from a mother plant with impunity, without endangering the mother plant.

#### 7.4.13.1.1 Cut Leaf Petioles

The process of the present invention generally proceeds as follows. Leaves are harvested from stock plants for which commercial propagation is desired. The petiole section of each leaf is sterilized with a soap-and-water wash followed by surface sterilization using a solution containing soap and hypochlorite bleach, or a sequence of ethanol and bleach rinses. A good sterilization protocol rinses the petiole tissue in 70% aqueous ethanol for 1 minute, followed by a 15 minute rinse with 10% aqueous bleach, followed by two rinses with sterile water.

After sterilization, the leaf petioles are cut into approximately 1 cm pieces. The cut leaf petioles are inoculated with *Actrobacterium* cells which contain the gene sequence desired to be transferred to the plant cells, preferably together with a marker gene such as the kanamycin resistance gene known in the art. The inoculation can be as simple as the physical mixing of the cut leaf petioles with the Agrobacterium cells, with an approximate 30 minute incubation at ambient room or greenhouse temperatures.

#### 7.4.13.1.2 Use of a Marker Gene

Those skilled in the art know the significance of the use of a marker gene, but it is instructive to review that technology here. If a genetic sequence to be transplanted includes both the gene (or antisense gene) of interest adjacent a marker gene such as an antibiotic-resistance gene, the successfully genetically transformed cells can easily be separated from any cells in which the desired transformation did not occur. As a practical matter in plant propagation, a number of explants or other regenerative plant cells can be exposed to the gene/marker gene combination and then screened for successful transformants by, for example, inducing and growing the plantlets in culture medium containing the antibiotic for which the marker gene imparts resistance. If any plant grows in the antibiotic -containing medium, it will also have been transformed with respect to the desired gene adjacent the antibiotic-resistance gene. Explants or other cells which may not have underwent genetic transformation merely die in the culture medium--due to antibiotic susceptibility--and disappear.



#### 7.4.13.1.3 Antisense Molecular Biology

Those skilled in the art also understand the significance of "antisense" molecular biology, but it should be borne in mind that primarily the present invention is intended to create transformants having antisense genes per se, and preferably not organisms containing vector-borne antisense mRNAs to prevent transcription of intact, or non-antisense, genes.

Transformation to create antisense genes is known in the art as exemplified by van der Krol, et al., "Antisense Chalcone Synthase Genes in Petunia Visualization of Variable Transgene Expression," Mol. Gen. Genet. (Molecular & General Genetics) Vol. 220, No. 2, pp. 204-212 (1990).

#### 7.4.13.1.4 Growth Period

##### 7.1.4.13.1.4.1 Transfer to Culture Medium

The inoculated petiole sections are then transferred to separate test tubes or vials containing culture medium. The culture medium contains vitamins, minerals, a food source and at least one growth regulator.

The food source usually includes the Murashige Skoog salt known in the art, and preferably also includes additional food/energy sources, most preferably fresh coconut milk, as well as Agrobacterium virulence enhancers such as acetosyringone. An essential growth regulator is a benzylamino compound chemically equivalent to the most preferred benzylaminopurineriboside or the benzylaminopurines generally. The use of this class of growth regulators gives unexpectedly good results over the use of other growth regulators such as 2,4- dichlorophenoxyacetic acid, kinetin, gibberellic acid, abscisic acid or 6- - dimethylallylaminopurine ( $N^6$  - [2- isopentenyl] adenine). Additional auxin and/or cytokinin growth regulators (indole acetic acid, indole butyric acid, benzylamine, benzyladenine, additional benzylaminopurine, alpha naphthylacetic acid and others) may also be present if they are in addition to, and not in substitution for, the benzyl/amino growth regulator selected.

The test tubes or vials are maintained for five days to two weeks in complete darkness, at a temperature of about 25° C. Over the five day to two week period, the section enlarges slightly and the ends form callus.

Miniature shoots start forming intermittently on the callused ends of the petiole section.

#### **7.4.13.1.4.2 Transfer to a Magenta Vial or Box Known in the Art**

After five days to two weeks, the enlarged petiole section bearing the miniature shoots is transferred from the test tube or vial to a Magenta vial or box known in the art. The enlarged petiole sections are housed five- to-a-Magenta vial. The same growing medium as was originally charged to the test tube or vial is likewise charged to the Magenta vial, and in any event coconut milk should be present in the culture medium at this stage of the process. Also added to the medium is kanamycin (assuming the Agrobacterium contained the kanamycin resistance gene) and carbenicillin to kill any excess Agrobacterium. The Magenta vials are then maintained, under the same conditions as were the test tubes or vials, for an additional five to eight weeks in the dark and at about 25° C. The Magenta vials are then exposed to 5-10 weeks of 16 hours of light daily, in which the temperature is maintained at 72° F with 690 foot candles (6900 lux) of cool fluorescent light. During this time the petiole sections grow into enlarged clumps; the shoots elongate and turn into plantlets and many more shoots form. Once plantlets appear, they are transferred to fresh media containing kanamycin, carbenicillin and no growth hormones.

#### **7.4.13.1.5 Rooting**

After the total growth period has elapsed, the clumps are removed and placed in sterile water. The individual plants are dissected out of the clump with a sterile scalpel. Each individual plant essentially has a series of leaves and nodes and is at least ½" high, but usually no roots are present. The individual plants are placed in RUBBER DIRT™ or other soil or soil-like growth media or growth media plugs, where rooting then takes place. Many varieties of *Pelargonium x domesticum* have been successfully tissue cultured through leaf petioles and multiplied, both with and without transgenic transformation via Agrobacterium. Morphologic variation has been minimal and within commercially

acceptable limits for finished plant material. Other plants may be propagated by this tissue culture technique/transgenic technique also.

#### **7.4.13.2 Creating an Agrobacterium Cell Containing the Desired Vector**

The creation of the Agrobacterium cell containing the desired vector can be accomplished by means known in the art. Structural and regulatory genes to be inserted may be obtained from depositories, such as the American Type Culture Collection, Rockville, MD, 20852, as well as by isolation from other organisms, typically by the screening of genomic or cDNA libraries using conventional hybridization techniques. Typical hybridization techniques are disclosed in Sambrook, et al., Molecular Cloning--Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989). Screening may be performed by (1) nucleic acid hybridization using homologous genes or heterologous genes from other organisms, (2) probes synthetically produced to hybridize to particular sequences coding for desired protein sequences, or (3) DNA sequencing and comparison to known sequences. Sequences for specific genes may be found in various computer databases, including GenBank, National Institutes of Health, or the database maintained by the United States Patent and Trademark Office.

The genes of interest may also be identified by antibody screening of expression libraries with antibodies made against homologous proteins to identify genes encoding for homologous functions. Transposon tagging can also be used to aid the isolation of a desired gene. Transposon tagging typically involves mutation of the target gene. A mutant gene is isolated in which a transposon has inserted into the target gene and altered the resulting phenotype.

##### **7.4.13.2.1 Isolation of the Mutated Gene**

Using a probe for the transposon, the mutated gene can be isolated. Then, using the DNA adjacent to the transposon in the isolated, mutated gene as a probe, the normal wild-type allele of the target gene can be isolated. Such techniques are taught, for example, in McLaughlin and Walbot, Genetics, Vol. 117 pp. 771-776 (1987), as well as numerous other references.

#### 7.4.13.2.2 Reporter Gene

In addition to the functional gene and the selectable marker gene, the DNA sequences may also contain a reporter gene which facilitates screening of the transformed shoots and plant material for the presence and expression of endogenous DNA sequences. Exemplary reporter genes include  $\beta$ -glucuronidase and luciferase.

#### 7.4.13.2.3 Transfer Regions of a Suitable Plasmid

As described above, the exogenous DNA sequences are introduced into the area of the explants by incubation with Agrobacterium cells which carry the sequences to be transferred within a transfer DNA (T-DNA) region found on a suitable plasmid, typically the Ti plasmid. Ti plasmids contain two regions essential for the transformation of plant cells. One of these, the T-DNA region, is transferred to the plant nuclei and induces tumor formation. The other, referred to as the virulence (vir) region, is essential for the transfer of the T-DNA but is not itself transferred. By inserting the DNA sequence to be transferred into the T-DNA region, introduction of the DNA sequences to the plant genome can be effected. Usually, the Ti plasmid will be modified to delete or to inactivate the tumor-causing genes so that they are suitable for use as a vector for the transfer of the gene constructs of the present invention. Other plasmids may be utilized in conjunction with Agrobacterium for transferring the DNA sequences of the present invention to the plant cells.

The construction of recombinant Ti plasmids may be accomplished using conventional recombinant DNA techniques, such as those described by Sambrook et al. (1989). Frequently, the plasmids will include additional selective marker genes which permit manipulation and construction of the plasmids in suitable hosts, typically bacterial hosts other than Agrobacterium, such as *E. coli*. In addition to the above-described kanamycin resistance marker gene, other exemplary genes are the tetracycline resistance gene and the ampicillin resistance gene, among others.

#### 7.4.13.2.4 Transcriptional and Translational Control Sequences

The genes within the DNA sequences will typically be linked to appropriate transcriptional and translational control sequences which are suitable for the *Pelargonium* plant host. For example, the gene will typically be situated at a distance from a promoter

corresponding to the distance at which the promoter is normally effective in order to ensure transcriptional activity. Usually, a polyadenylation site and transcription termination site will be provided at the 3'-end of the gene coding sequence. Frequently, the necessary control functions can be obtained together with the structural gene when it is isolated from a target plant or other host. Such intact genes will usually include coding sequences, intron (s), a promoter, enhancers, and all other regulatory elements either upstream (5') or downstream (3') of the coding sequences.

#### 7.4.13.2.5 Using the Binary Vector System to Introduce the DNA Sequence

The binary vector system generally discussed above may be used to introduce the DNA sequence according to the present invention. A first plasmid vector strain would carry the T-DNA sequence while a second plasmid vector carries a virulence (vir) region. By incubating Agrobacterium cells carrying both plasmids with the explant, infection of the plant material is thus achieved.

#### 7.4.13.2.6 The T-DNA Plasmid

Any one of a number of T-DNA plasmids can be used with such a binary vector system, the only requirement being that one be able to select independently for the two plasmids. The T-DNA plasmid in a preferred embodiment comprises a heterologous promoter which promotes the transcription of one or more genes within the exogenous DNA fragment (s). An example is the Cauliflower Mosaic Virus 35S promoter (Odell et al., Nature, Vol. 313, pp. 810-812 (1985) among others.

#### 7.4.13.2.7 Agrobacterium Species

Suitable Agrobacterium species include Agrobacterium tumefaciens and Agrobacterium rhizogenes. While the wild-type Agrobacterium rhizogenes may be used, the Agrobacterium tumefaciens should be disarmed by deactivating its tumor activating capacity prior to use.

Preferred Agrobacterium tumefaciens strains include LBA4404, as described by Hoekema et al., Nature, Vol. 303, pp. 179-180 (1983) and EHA101 as described by Hood et al., J. Bacteriol., Vol. 168, pp. 1291-1301 (1986). A preferred Agrobacterium

rhizocrenes strain is 15834, as described by Birot et al., Plant Physiol. Biochem., Vol. 25, pp. 323-325 (1987).

#### 7.4.13.2.8 Antisense Sequences

As the Agrobacterium strains carrying the desired exogenous DNA sequences are being prepared, the following antisense sequences are preferred for use in the present process, and are particularly preferred in transforming regal Pelargonium petiole explants according to the present method. These sequences are the known sequences for ACC Synthase (1-aminocyclopropane-1-carboxylate synthase) and ACC Oxidase (1-aminocyclopropane-1-carboxylate oxidase), reversed to create antisense sequences.

#### 7.4.13.2.9 Culturing the Agrobacterium Strains

After the Agrobacterium strains carrying the desired exogenous DNA sequences have been prepared, they will usually be cultured for a period of time prior to incubation with the explant material. Initially, the Agrobacterium may be cultured on a solid media including nutrients, an energy source and a gelling agent. Suitable nutrients include salts, tryptone and yeast extracts, while most sugars are suitable as the energy source and the gelling agent can be TC agar or bactoagar or other similar products. The Agrobacterium cells are typically cultured for about two to five days, preferably in the dark at about 23-28° C, and are collected by scraping before browning (while still a white color). The cells are scraped from the medium and suspended in a liquid medium such as L-broth, pH 6.9-7.1, preferably 7.0. The bacteria are cultured in liquid medium for 8-36 hours, preferably 12-20, on a shaker (50-220 rpm, preferably 100-120 rpm) at 23-28° C. At the end of this period the bacteria are diluted to an optical density of 0.3 and cultured for 2-6, preferably 4, hours (on a shaker, 23-28° C). The Agrobacterium cells thus incubated are ready for inoculation onto explant material such as Pelargonium petiole sections.

When Agrobacterium cells are inoculated onto Pelargonium x domesticum petiole explants, and after the coincubation discussed above, the tissue culture method will proceed generally in accordance with the disclosure of U.S. Application Serial No. 08/149,702 filed November 9, 1993, which is a Continuation of U.S. Application Serial No. 07/690,073 filed April 23, 1991, by Wendy Oglevee- O'Donovan and Eleanor Stoots, both of which applications are hereby incorporated herein by reference. By means known

in the art, carbenicillin can be added to the coincubated petiole/Agrobacterium cells to kill excess Agrobacterium cells, after transfection has taken place.

#### 7.4.13.2.10 Selecting an Antibiotic

The key is to select an antibiotic which will kill the Agrobacterium without harming the explant material. An additional amount of the same antibiotic may be provided in the ensuing tissue culture method, to assure final removal of any viable Agrobacterium cells.

#### 7.4.13.2.11 Confirming Transformation

After green transformed shoots are approximately 1/2" tall, they can then be transplanted to soil within a greenhouse or elsewhere in a conventional manner for tissue culture plantlets. Transformation of the resulting plantlets can be confirmed by assaying activity for the selection marker, or by assaying the plant material for any of the phenotypes which have been introduced by the exogenous DNA. Suitable assay techniques include polymerase chain reaction (PCR), restriction enzyme digestion, Southern blot hybridization and Northern blot hybridization.

#### 7.4.13.3 Commercial Production

The present invention represents a breakthrough in the commercial production and genetically transformed plants. Because the method uses petiole tissue from a grower's mother plant (a stock plant), the starting petiole explants have a commercially desirable morphology to begin with--by definition. However, if the mother plant could be improved by genetic transformation of some type, for example to deactivate a gene which expresses an enzyme in the ethylene synthesis pathway, the progeny of the mother plant may thus be improved in this one way over their parent stock. The petiole tissue from the stock plant, plus the genetic transformation from the Agrobacterium, yield both an improved genetic makeup of the commercially produced plants--although with preserved desired morphology from the mother plant--and at the same time the high yields possible only with the generation of many plantlets in a single generation's growth in tissue culture. In summary, with the present method a single genetically transformed mother plant can yield literally thousands of offspring plants. No one in the prior art has attempted to combine these two previously disparate technologies to achieve a unique method in which the result

is no less than a commercially viable technique for making genetically recombinant plants in commercially feasible numbers.



#### **7.4.14 TRANSFORMATION OF PLANT CELLS USING ALTERNATIVE METHODS**

A variety of techniques are available for the introduction of the genetic material into or transformation of the plant cell host. However, the particular manner of introduction of the plant vector into the host is not critical to the practice of the present invention, and any method which provides for efficient transformation may be employed. In addition to transformation using plant transformation vectors derived from the tumor-inducing (Ti) or root-inducing (Ri) plasmids of Agrobacterium, alternative methods could be used to insert the DNA constructs of the present invention into plant cells. Such methods may include, for example, the use of liposomes, transformation using viruses or pollen, chemicals that increase the direct uptake of DNA (Paszkowski et al., EMBO J., 3, 2717 (1984)), microinjection (Crossway et al., Mol. Gen. Genet., 202, 179 (1985)), electroporation (Fromm et al., Proc. Natl. Acad. Sci. US , 82, 824 (1985)), or high-velocity microprojectiles (Klein et al., Nature, 327, 70 (1987)).

##### **7.4.14.1 Plant Tissue Source or Cultured Plant Cell**

The choice of plant tissue source or cultured plant cells for transformation will depend on the nature of the host plant and the - transformation protocol. Useful tissue sources include callus, suspension culture cells, protoplasts, leaf segments, stem segments, tassels, pollen, embryos, hypocotyls, tuber segments, meristematic regions, and the like.

The tissue source is regenerable, in that it will retain the ability to regenerate whole, fertile plants following transformation.

##### **7.4.14.2 Conditions During Transformation**

The transformation is carried out under conditions directed to the plant tissue of choice. The plant cells or tissue are exposed to the DNA carrying the present multi-gene expression cassette for an effective period of time. This may range from a less-than-one-second pulse of electricity for electroporation, to a two-to-three day co-cultivation in the presence of plasmid-beazing Agrobacterium cells. Buffers and media used will also vary with the plant tissue source and transformation protocol. Many transformation protocols employ a feeder layer of suspended culture cells (tobacco or Black Mexican Sweet Corn,

for example) on the surface of solid media plates, separated by a sterile filter paper disk from the plant cells or tissues being transformed.

Following treatment with DNA, the plant cells or tissue may be cultivated for varying lengths of time prior to selection, or may be immediately exposed to a selective agent such as those described hereinabove.

#### **7.4.14.3 Inhibitory Agent**

Protocols involving exposure to Agrobacterium will also include an agent inhibitory to the growth of the Agrobacterium cells. Commonly used compounds are antibiotics such as cefotaxime and carbenicillin. The media used in the selection may be formulated to maintain transformed callus or suspension culture cells in an undifferentiated state, or to allow production of shoots from callus, leaf or stem segments, tuber disks, and the like.

#### **7.4.15 METHOD FOR TRANSFORMATION OF THE TARGET PLANT**

The methods used for the actual transformation of the target plant are not critical to this invention. The transformation of the plant is preferably permanent, e.g. by integration of introduced sequences into the plant genome, so that the introduced sequences are passed onto successive plant generations. There are many plant transformation techniques well-known to workers in the art, and new techniques are continually becoming known. Any technique that is suitable for the target plant can be employed with this invention. For example, the sequences can be introduced in a variety of forms, such as a strand of DNA, in a plasmid, or in an artificial chromosome, to name a few. The introduction of the sequences into the target plant cells can be accomplished by a variety of techniques, as well, such as calcium phosphate-DNA co-precipitation, electroporation, microinjection, Agrobacterium infection, liposomes or microprojectile transformation. Those of ordinary skill in the art can refer to the literature for details, and select suitable techniques without undue experimentation.

##### **7.4.15.1 Introduction of Sequences into Target Plant Cells**

It is possible to introduce the recombinase gene, in particular, into the transgenic plant in a number of ways. The gene can be introduced along with all of the other basic sequences, as in the first preferred embodiment described above. The repressible promoter/recombinase construct can be also introduced directly via a viral vector into a transgenic plant that contains the other sequence components of the system. Still another method of introducing all the necessary sequences into a single plant is the second preferred embodiment described above, involving a first transgenic plant containing the transiently-active promoter/structural gene sequences and the blocking sequence, and a second transgenic plant containing the recombinase gene linked to a germination-specific plant-active promoter, the two plants being hybridized by conventional to produce hybrid progeny containing all the necessary sequences.

It is also possible to introduce the recombinase itself directly into a transgenic plant as a conjugate with a compound such as biotin, that is transported into the cell. See Horn, et al. (1990).

##### **7.4.15.2 Direct or Vectored Transformation**

Various methods for direct or vectored transformation of plant cells, e. g., plant protoplast cells, have been described, e.g., in above-cited PCT application WO 95/14099. As noted in that reference, promoters directing expression of selectable markers used for plant transformation (e.g., nptII) should operate effectively in plant hosts. One such promoter is the nos promoter from native Ti plasmids (Heffera-Estrella, et al., Nature 303:209-213 (1983). Others include the 35S and 19S promoters of cauliflower mosaic virus (Odell, et al., Nature 313:810-812 (1985) and the 2' promoter (Velten, et al., EMBO J. 3:2723-2730 (1984).

In one preferred embodiment, the embryo and endosperm. of mature seeds are removed to exposed scutulum tissue cells. The cells may be transformed by DNA bombardment or injection, or by vectored transformation, e.g., by Agrobacteriwn infection after bombarding the scuteller cells with microparticles to make them susceptible to Agrobacteriwn infection (Bidney et al., Plant Mol. Biol. 18:301-313, 1992).

One preferred transformation follows the methods detailed generally in Sivamani, E. et al., Plant Cell Reports 15:465 (1996); Zhang, S., et al., Plant Cell Reports 15:465 (1996); and Li, L., et al., Plant Cell Reports 12:250 (1993). Briefly, rice seeds are sterilized by standard methods, and callus induction from the seeds is carried out on MB media with 2,41). During a first incubation period, callus tissue forms around the embryo of the seed. By the end of the incubation period, (e.g., 14 days at 28°C) the calli are about 0.25 to 0.5 cm in diameter. Callus mass is then detached from the seed, and placed on fresh NB media, and incubated again for about 14 days at 28°C. After the second incubation period, satellite calli developed around the original "mother" callus mass.

These satellite calli were slightly smaller, more compact and defined than the original tissue. It was these calli were transferred to fresh media. The "mother " calli was not transferred. The goal was to select only the strongest, most vigorous growing tissue for fiwffier culture.

Calli to be bombarded are selected from 14-day-old subcultures. The size, shape, color and density are all important in selecting calli in the optimal physiological condition

for transformation. The calli should be between .8 and 1.1 mm in diameter. The calli should appear as spherical masses with a rough exterior.

Transformation is by particle bombardment, as detailed in the references cited above. After the transformation steps, the cells are typically grown under conditions that permit expression of the selectable marker gene. In a preferred embodiment, the selectable marker gene is HPH. It is preferred to culture the transformed cells under multiple rounds of selection to produce a uniformly stable transformed cell line.

#### **7.4.16 SUBCULTURING CELLS OR CALLUS GROWING IN NORMALLY INHIBITORY CONCENTRATIONS OF THE SELECTIVE AGENTS**

Cells or callus observed to be growing in the presence of normally inhibitory concentrations of the selective agents are presumed to be transformed and may be subcultured several additional times on the same medium to remove non-resistant sections. The cells or calli can then be assayed for the presence of the viral gene cassette, or may be subjected to known plant regeneration protocols. In protocols involving the direct production of shoots, those shoots appearing on the selective media are presumed to be transformed and may be excised and rooted, either on selective medium suitable for the production of roots, or by simply dipping the excised shoot in a root-inducing compound and directly planting it in vermiculite.

#### 7.4.17 SELECTING FOR MULTI-VIRAL RESISTANCE

In order to produce transgenic plants exhibiting multi-viral resistance, the viral genes must be taken up into the plant cell and stably integrated within the plant genome. Plant cells and tissues selected for their resistance to an inhibitory agent are presumed to have acquired the selectable marker gene encoding this resistance during the transformation treatment.

Since the marker gene is commonly linked to the viral genes, it can be assumed that the viral genes have similarly been acquired. Southern blot hybridization analysis using a probe specific to the viral genes can then be used to confirm that the foreign genes have been taken up and integrated into the genome of the plant cell. This technique may also give some indication of the number of copies of the gene that have been incorporated. Successful transcription of the foreign gene into mRNA can likewise be assayed using Northern blot hybridization analysis of total cellular RNA and/or cellular RNA that has been enriched in a polyadenylated region. mRNA molecules encompassed within the scope of the invention are those which contain viral specific sequences derived from the viral genes present in the transformed vector which are of the same polarity to that of the viral genomic RNA such that they are capable of base pairing with viral specific RNA of the opposite polarity to that of viral genomic RNA under conditions described in Chapter 7 of Sambrook et al. (1989). mRNA molecules also encompassed within the scope of the invention are those which contain viral specific sequences derived from the viral genes present in the transformed vector which are of the opposite polarity to that of the viral genomic RNA such that they are capable of base pairing with viral genomic RNA under conditions described in Chapter 7 of Sambrook et al. (1989).

The presence of a viral gene can also be detected by immunological assays, such as the double-antibody sandwich assays described by Namba, et al., *Gene*, 107, 181 (1991) as modified by Clark et al., *J. Gen. Virol.*, 34, 475 (1979). See also, Namba et al., *Phytopathology*, 82, 940 (1992).

Virus resistance can be assayed via infectivity studies as generally disclosed by Namba et al., *ibid.*, wherein plants are scored as symptomatic when any inoculated leaf shows veinclearing, mosaic or necrotic symptoms.

It is understood that the invention is operable when either sense or anti-sense viral specific RNA is transcribed from the expression cassettes described above. That is, there is no specific molecular mechanism attributed to the desired phenotype and/or genotype exhibited by the transgenic plants.

Thus, protection against viral challenge can occur by any one or any number of mechanisms.

It is also understood that virus resistance can occur by the expression of any virally encoded gene. Thus, transgenic plants expressing a coat protein gene or a non-coat protein gene can be resistant to challenge with a homologous or heterologous virus..



#### **7.4.18 CELL CULTURE PRODUCTION OF MATURE HETEROLOGOUS PROTEIN**

Transgenic cells, typically callus cells, are cultured under conditions that favor plant cell growth, until the cells reach a desired cell density, then under conditions that favor expression of the mature protein under the control of the given promoter. Preferred culture conditions are described herein. Purification of the mature protein secreted into the medium is by standard techniques known by those of skill in the art.

In one embodiment of the invention, in which BPN' is secreted as the proBPN' form of the enzyme, the chaperon "pro" moiety of the enzyme facilitates enzyme folding and is cleaved from the enzyme, leaving the active mature form of BPN'. In another embodiment, the mature enzyme is co-expressed and co-secreted with the "pro" chaperon moiety, with conversion of the enzyme to active form occurring in presence of the free chaperon (Eder et al., Biochem. (1993) L2:18-26; Eder et al, (1993) J. Mol. Biol. 223:293-304). In yet another embodiment of the invention, the BPN' is secreted in inactive form at a pH that may be in the 6-8 range, with subsequent activation of the inactive form, e.g., after enzyme isolation, by exposure to the "pro" chaperon moiety, e.g., immobilized to a solid support. In both of these embodiments, the culture medium is maintained at a pH of between 5 and 6, preferably about 5.5 during the period of active expression and secretion of BPN', to keep the BPN', which is normally active at alkaline pH, at a pH below optimal activity.

##### **7.4.18.1 Production of Mature Heterologous Protein in Germinatin Seeds**

In this embodiment, monocot cells transformed as above are used to regenerate plants, seeds from the plants are harvested and then germinated, and the mature protein is isolated from the germinated seeds.

Plant regeneration from cultured protoplasts or callus tissue is carried by standard methods, e.g., as described in Evans et al., HANDBOOK OF PLANT CELL CULTURE Vol. 1: (MacMillan Publishing Co. New York, 1983); and Vasil I.R. (ed.), CELL CULTURE AND SOMATIC CELL GENETICS OF PLANTS, Acad. Press, Orlando, Vol. 1, 1984, and Vol. 111, 1986, and as described in the above-cited PCT application.

#### 7.4.18.2 Seed Germination Conditions

The transgenic seeds obtained from the regenerated plants are harvested, and prepared for germination by an initial steeping step, in which the seeds immersed in or sprayed with water to increase the moisture content of the seed to between 35-45%. This initiates germination. Steeping- typically takes place in a steep tank which is typically fitted with a conical end to allow the seed to flow freely out. The addition of compressed air to oxygenate the steeping process is an option.

The temperature is controlled at approximately 22-C depending on the seed.

After steeping, the seeds are transferred to a germination compartment which contains air saturated with water and is under controlled temperature and air flows. The typical temperatures are between 12-25-C and germination is permitted to continue for from 3 to 7 days.

Where the heterologous protein coding gene is operably linked to a inducible promoter requiring a metabolite such as sugar or plant hormone, e.g., 2 to 100 M gibberellic acid, this metabolite is added, removed or depleted from the steeping water medium and/or is added to the water saturated air used during germination. The seed absorbs the aqueous medium and begins to germinate, expressing the heterologous protein. The medium may then be withdrawn and the malting begun, by maintaining the seeds in a moist temperature controlled aerated environment. In this way, the seeds may begin growth prior to expression, so that the expressed product is less likely to be partially degraded or denatured during the process.

More specifically, the temperature during the imbibition or steeping phase will be maintained in the range of about 15-25°C, while the temperature during the germination will usually be about 20°C. The time for the imbibition will usually be from about 1 to 4 days, while the germination time will usually be an additional 1 to 10 days, more usually 3 to 7 days. Usually, the time for the malting does not exceed about ten days. The period for the malting can be reduced by using plant hormones during the imbibition, particularly gibberellic acid.

To achieve maximum production of recombinant protein from malting, the malting procedure may be modified to accommodate de-hulled and de-embryonated seeds, as described in above-cited PCT application WO 95/14099. In the absence of sugars from the endosperm, there is expected to be a 5 to 10 fold increase in RAmy3D promoter activity and thus expression of heterologous protein. Alternatively when embryoless half-seeds are incubated in 10 mM CaCl<sub>2</sub> and 5 M gibberellic acid, there is a 50 fold increase in RAmy1A promoter activity.

#### **7.4.18.3 Production of mature HAS**

Following the germination conditions as outlined above and further detailed herein, supernatant was analyzed by Western blot. Western blot analysis shows production of HSA in germinating rice seeds, with seed samples taken 24, 72, and 120 hours after induction with gibberellin. HSA production was highest approximately 24 hours post-induction. Bilirubin binding, a measure of correct folding of plant-produced HSA, is assayed according to the method presented herein.

#### **7.4.19 PRODUCTION OF MATURE HETEROLOGOUS PROTEIN IN MATURING SEEDS**

In this embodiment, monocot cells transformed as above are used to regenerate plants, and seeds from the plants are allowed to mature, typically in the field, with consequent production of heterologous protein in the seeds.

Following seed maturation, the seeds and their heterologous proteins may be used directly, that is, without protein isolation, where for example, the heterologous protein is intended to confer a benefit on the seed as a whole, for example, to enrich the seed in the selected protein.

Alternatively, the seeds may be fractionated by standard methods to obtain the heterologous protein in enriched or purified form. In one general approach, the seed is first milled, then suspended in a suitable extraction medium, e.g., an aqueous or an organic solvent, to extract the protein or metabolite of interest. If desired the heterologous protein can be further fractionated and purified, using standard purification methods.

The following examples are provided by way of illustration only and not by way of limitation. Those of skill will readily recognize a variety of noncritical parameters which could be changed or modified to yield essentially similar results.

#### **7.4.20 REGENERATION OF THE TRANSFORMED PLANT CELLS**

After transformation, the transformed plant cells must be regenerated.

The methods used to regenerate transformed cells into whole plants are not critical to this invention, and any method suitable for the target plant can be employed. The literature describes numerous techniques for regenerating specific plant types, (e.g., via somatic embryogenesis, Umbeck, et al., 1987) and more are continually becoming known. Those of ordinary skill in the art can refer to the literature for details and select suitable techniques without undue experimentation.

Plant regeneration from cultured protoplasts is described in Evans et al., Handbook of Plant Cell Cultures, Vol. 1: (MacMillan Publishing Co., New York, 1983); and Vasil I.R. (ed.), Cell Culture and Somatic Cell Genetics of Plants, Acad. Press, Orlando, Vol. I, 1984, and Vol.-III (1986), which are hereby incorporated is by reference.

It is known that practically all plants can be regenerated from cultured cells or tissues, including but not limited to, all major species of sugarcane, sugar beets, cotton, fruit trees, and legumes.

Means for regeneration vary from species to species of plants, but generally a suspension of transformed protoplasts or a petri plate containing explants is first provided. Callus tissue is formed and shoots may be induced from callus and subsequently rooted. Alternatively, embryo formation can be induced in the callus tissue. These embryos germinate as natural embryos to form plants. The culture media will generally contain various amino acids and hormones, such as auxin and cytokinins.. It is also advantageous to add glutamic acid and proline to the medium, especially for such species as corn and alfalfa. Efficient regeneration will depend on the medium, on the genotype, and on the history of the culture. If these three variables are controlled, then regeneration is usually reproducible and repeatable.

#### 7.4.21 BREEDING TECHNIQUES

After the expression cassette is stably incorporated in transgenic plants, it can be transferred to other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

Seed from plants regenerated from tissue culture is grown in the field and self-pollinated to generate true breeding plants. The progeny from these plants become true breeding lines which are evaluated for viral resistance in the field under a range of environmental conditions. The commercial value of viral-resistant plants is greatest if many different hybrid combinations with resistance are available for sale. The farmer typically grows more than one kind of hybrid based on such differences as maturity, disease and insect resistance, color or other agronomic traits. Additionally, hybrids adapted to one part of a country are not adapted to another part because of differences in such traits as maturity, disease and insect tolerance, or public demand for specific varieties in given geographic locations.

Because of this, it is necessary to breed viral resistance into a large number of parental lines so that many hybrid combinations can be produced.

Adding viral resistance to agronomically elite lines is most efficiently accomplished when the genetic control of viral resistance is understood. This requires crossing resistant and sensitive plants and studying the pattern of inheritance in segregating generations to ascertain whether the trait is expressed as dominant or recessive, the number of genes involved, and any possible interaction between genes if more than one are required for expression. With respect to transgenic plants of the type disclosed herein, the transgenes exhibit dominant, single gene Mendelian behavior. This genetic analysis can be part of the initial efforts to convert agronomically elite, yet sensitive lines to resistant lines. A conversion process (backcrossing) is carried out by crossing the original resistant line with a sensitive elite line and crossing the progeny back to the sensitive parent. The progeny from this cross will segregate such that some plants carry the resistance gene(s) whereas some do not. Plants carrying the resistance gene(s) will be crossed again to the sensitive parent resulting in progeny which segregate for resistance and sensitivity once more. This is repeated until the original sensitive parent has been

converted to a resistant line, yet possesses all of the other important attributes originally found in the sensitive parent. A separate backcrossing program is implemented for every sensitive elite line that is to be converted to a virus resistant line.

Subsequent to the backcrossing, the new resistant lines and the appropriate combinations of lines which make good commercial hybrids are evaluated for viral resistance, as well as for a battery of important agronomic traits. Resistant lines and hybrids are produced which are true to type of the original sensitive lines and hybrids. This requires evaluation under a range of environmental conditions under which the lines or hybrids will be grown commercially. Parental lines of hybrids that perform satisfactorily are increased and utilized for hybrid production using standard hybrid production practices.

#### **7.4.21.1 USE OF CONVENTIONAL CULTIVATION**

Once transgenic plants of this type are produced, the plants themselves can be cultivated in accordance with conventional procedure so that the DNA construct is present in the resulting plants. Alternatively, transgenic seeds are recovered from the transgenic plants. These seeds can then be planted in the soil and cultivated using conventional procedures to produce transgenic plants.

#### **7.4.21.2 PLANT VARIETIES**

The present invention can be used to make a variety of transgenic plants. The method is particularly suited for use with plants that are planted as a yearly crop from seed. These include, but are not limited to, fiber crops such as cotton and flax; dicotyledonous seed crops such as soybean, sunflower and peanut; annual ornamental flowers; monocotyledonous grain crops such as maize, wheat and sorghum; leaf crops such as tobacco; vegetable crops such as lettuce, carrot, broccoli, cabbage and cauliflower; and fruit crops such as tomato, zucchini, watermelon, cantaloupe and pumpkin.

The present invention can be utilized in conjunction with a wide variety of plants or their seeds.

Suitable plants include dicots and monocots. More particularly, useful crop plants can include: alfalfa, rice, wheat, barley, rye, cotton, sunflower, peanut, corn, potato, sweet potato, bean, pea, chicory, lettuce, endive, cabbage, brussel sprout, beet, parsnip, turnip, cauliflower, broccoli, turnip, radish, spinach, onion, garlic, eggplant, pepper, celery, carrot, squash, pumpkin, zucchini, cucumber, apple, pear, melon, citrus, strawberry, grape, raspberry, pineapple, soybean, tobacco, tomato, sorghum, papaya, and sugarcane.

Examples of suitable ornamental plants are: *Arabidopsis thaliana*, *Saintpaulia*, *petunia*, *pelargonium*, *poinsettia*, *chrysanthemum*, *carnation*, and *zinnia*.

The plants used in the process of the present invention are derived from monocots, particularly the members of the taxonomic family known as the Gramineae. This family includes all members of the grass family of which the edible varieties are known as cereals. The cereals include a wide variety of species such as wheat (*Triticwn* sps.), rice (*Oryza* sps. ) barley (*Hordewn* sps.) oats, (*Avena* sps.) rye (*Secale* sps.), corn (*Zea* sps.) and millet (*Pennisettum* sps.). In the present invention, preferred family members are rice and barley.



#### **7.4.22 IDENTIFICATION AND LOCALIZATION AND INTROGRESSION INTO PLANTS OF DESIRED MULTIGENIC TRAITS WITH RFLP TECHNOLOGY**

The invention typically involves genetic linkage maps constructed with RFLP technology and the use of RFLP probes to correlate those probes with Quantitative Trait Loci (QTL) and the degree of inheritance of particular multigenic traits.

##### **7.4.22.1 Determining the Degree of Inheritance of Particular Multigenic Traits**

In one embodiment, a plant source (designated  $P_1$ ) having a desired multigenic trait--for example, increased height--is recovered and crossed with a second plant (designated  $P_2$ ) having essentially or substantially opposite characteristics, that is, decreased height. Heterozygote plants from the  $F_1$  population are selfed to create a segregating ( $F_2$ ) plant population which exhibit a gradient with respect to height, i.e., with respect to the degree of expression of the multigenic or quantitative trait of interest.

##### **7.4.22.1.1 RFLP Probe Tests**

Quantitative values for the trait of interest (height) are determined and assigned to each individual parent plant,  $F_1$  population plant, and  $F_2$  segregation plant and a genomic DNA sample from each plant is prepared for Southern blotting. Following preparation for a Southern blot--which may be constructed to contain, for example, DNA from 25 to 50 or more different  $F_2$  plants--an RFLP probe is randomly chosen or selected from an RFLP genetic linkage map and hybridized to create the blot. Additional Southern blots are constructed using other RFLP probes. As indicated above, the RFLPs to be used for this purpose, i.e., the indirect selection of a QTL, may but need not be randomly chosen. They can be selected in systematic fashion from the RFLP genetic linkage map. For example, for a trait of completely unknown location, several spaced RFLPs from each of the 10 chromosomes of maize may be selected for Southern blot testing for the location of DNA associated with a desired height. Alternatively, of course, all mapped RFLPs may be used.

Following these Southern blot constructs, a matrix may be prepared having an identification of each plant that has been tested, followed by its quantitative trait measurement (height) and the genotype as revealed by each RFLP probe tested. Typically, only three genotypes will be seen:  $P_1$ ,  $P_2$  and  $F_1$ , the latter being heterozygous and having one chromosome from each parent. Thus, from the matrix all plants can be grouped into

one of three RFLP genotypic categories:  $P_1P_1$ ,  $P_1P_2$  or  $P_2P_2$ . If, with one or more RFLPs, the so-grouped plants, when averaged, all show approximately equal expression of the trait of interest, i.e., the average height of plants in all groups is about the same, that RFLP is deemed noninformative. In other words, there was no association between the trait of interest and that particular RFLP or RFLPs. The genotype of the plant at the location of the RFLP was not relevant to the trait of interest.

#### **7.4.22.1.2 Alteration of Experiment**

Another RFLP, however, may show association with height. For example, the average height of the maize plants, respectively, in each of the  $P_1P_1$ ,  $P_1P_2$  and  $P_2P_2$  groups for a different RFLP, respectively, may be 3 feet, 4 feet and 5 feet. With this information, it may be presumed that this RFLP, as revealed by the degree of its correlation to the  $P_2P_2$  genotype, hybridizes to maize DNA in the area of a gene for height.

#### **7.4.22.1.3 Use of a Genetic Linkage Map**

In the above-described manner, it is possible to review results from a first group of RFLP probes used to screen for association to the trait of interest. Use of an RFLP genetic linkage map allows the selection of further RFLPs to be tested on an objective, rather than random, basis. Correlation may be improved by testing RFLPs located on either side of that RFLP or RFLPs which initially showed the strongest association. Once the best probe or probes are identified they may then be utilized, by way of example, in a breeding program to select plants having a desired height.

It is to be noted, of course, that in a multigenic system, there may be three, four, or more different genes contributing to one trait. In such a situation there may, therefore, be many different quantitative expressions of that trait and no one gene can account for, or be relied upon to predict, that expression. We have further determined that the relative importance of each correlating RFLP can be determined. Particular values can be assigned to those RFLPs and utilized in a mathematical model to assist in predicting the degree of trait expression in a particular plant.

#### **7.4.22.1.4 Analysis of Variance**

In the case of RFLP analysis, there are three different genotypic classes, i.e., AA, Aa and aa. Among three classes there are two possible contrasts and, hence, there are 2 degrees of freedom among genotypic classes. Moreover, the two possible contrasts among genotypes can be specifically partitioned into linear and quadratic sources, each with one degree of freedom. The linear contrast would have the values -1, 0 and +1 for the three genotypic classes AA, Aa and aa, respectively. This effectively compares the difference between the AA class and the aa class, i.e., the parental classes. The values for the quadratic contrast are 1, -2 and 1 for the three genotypic classes and, therefore, compare the difference between the mean of the parental AA and aa classes to the heterozygote class Aa.

The error, or deviation from regression, reflects the failure of the observed points to be exactly on the regression line. In the case of RFLP "A" we can observe that the mean square for the linear contrast is approximately 100 times larger than that of the quadratic contrast and about 10 times larger than the error mean square. The magnitudes of the mean squares from the analysis of variance indicates that linear regression can account for the majority of the variance among the plant heights. The  $r^2$  (coefficient of determination) value can be defined as the sums of the linear and quadratic mean squares divided by the total phenotypic variance (the total sums of squares of deviations from mean). The  $r^2$  value for "A" indicates that 89% of the observed variance for plant height can be explained as due to differences among RFLP "A" genotypic classes. Similarly, the analysis of variance for RFLP "B" indicates that none of the variance for plant height can be explained due to differences among genotypic classes. The analysis of variance for RFLP "C" indicates that 53% of the variance can be explained due to differences among genotypic classes and that both the linear and quadratic contrast are important. Thus, the relative magnitudes of mean squares from the analysis of variance provides a description of the relationship between RFLPs and plant height. The conclusions coincide with those obtained from observation of the plots of plant height by genotypic classes.

The means for each genotypic class are also presented in Table 2. These means represent the average phenotypic performance for each of the 3 genotypic classes. For RFLP "A", the mean of the parent 1 class (aa) is 12.5, the mean for the heterozygous class (Aa) is 62.5, and the mean for the homozygous parent 2 class (AA) is 100 plant height

units. The linear regression line slope ( $b_1$ ) is also presented for each RFLP. Thus, the linear regression of plant height on genotypic classes for RFLP "A" is 45. This, as noted above, can be interpreted as indicating that each time an "A" allele is substituted for an a allele that plant height will increase 45 units. The slope thus permits comparisons among RFLPs and helps to identify those RFLPs most strongly associated with the expression of a trait of interest.

Once the RFLPs most strongly associated with the expression of a trait are identified, the effects of each RFLP may be combined into a multiple regression model which will permit prediction of expression of the trait of interest based on knowledge of the genotypes of specific RFLP. The effects of each RFLP are not strictly additive because the effects of RFLP may be correlated. For example, in this case, RFLP "A" and "C" may each provide some unique information, but part of the information provided by one RFLP may also be provided by another. Thus, the average effects of allelic substitutions may not be simply added together to provide a predictive model.

#### 7.4.22.2 General Form of Multiple Regression Model

The general form of a useful multiple regression predictive model is shown below:

$$y = \mu + b_1 (\text{genotype of RFLP locus 1}) + b_2 (\text{genotype of RFLP locus 2})$$

where  $y$  is the predicted expression of a trait,  $\mu$  is the weighted mean expression of the trait for the population,  $b_1$  is the coefficient associated with a specific RFLP and so on. The genotype of a RFLP will be -1, 0 or 1 for genotypic classes AA, Aa and aa if the RFLP is additive (linear), or 1, -2 or 1 if the RFLP is quadratic (non-additive).

Determination of the "best" multiple regression model requires an iterative process of substitution of RFLPs into and out of the model and the evaluation of interaction (epistatic) effects of RFLPs. For example, if two independent genes act together to provide expression of a trait, the genotypic class may involve the products of linear and/or quadratic genotypic values. The process for determination of the "best" multiple regression model can be done using stepwise regression procedures or by comparison of partial and sequential sums of squares from the regression models.

#### 7.1.4.22.2.1 Example of Model Building

For example, in the present hypothetical, the logical RFLPs to include in the model would be the linear contrast for RFLP "A" and both the linear and quadratic contrasts for RFLP "C". The partial and sequential sums of squares (SS) for a model including RFLP A and C are the following:

Source	sequential SS	partial SS
linear "A"	5580	2812
linear "C"	62	236
quadratic "C"	503	503

The magnitudes of the sums of squares indicates that linear "A" is the most important determinate in explaining the variance for plant height. However, the reduction in partial SS (2812) vs. sequential SS (5580) for linear "A" suggests that much of the effect of linear "A" is accounted for by the linear and quadratic effects of "C". It should be noted that if the effects of each locus were completely independent there would be no differences in the partial and sequential sums of squares. Changes in the magnitudes of these sums of squares in an analysis of variance indicates that the effects of the different RFLP are correlated.

#### 7.4.22.2.2 Another Model Constructed (Eliminate Linear Contrast)

The iterative process of substituting linear and quadratic contrasts for different RFLPs can continue until a final predictive model is constructed. The objective of the model is to maximize the  $r. sup.2$  (coefficient of determination) value using as few RFLP as possible as predictive variables. The final model will generally eliminate RFLP which might flank a particular gene of interest because both RFLP will typically be contributing the same information. In addition, the final model might contain effects with reflect interactions between RFLPs.

In the present hypothetical, the next step in the process of model building might be to eliminate the linear contrast for RFLP "C". The results of this model are the following:

Source	sequential SS	partial SS
--------	---------------	------------

linear "A"	5580	5127
quadratic "C"	328	328

#### 7.4.22.3 Final Mathematical Model

The general agreement in magnitude between partial and sequential SS suggest that the effects of linear "A" and quadratic "C" are relatively independent. The  $r_{\text{sup.2}}$  value for this model is 0.95. Thus, 95% of the variance for plant height can be explained as due to differences among genotypic classes at RFLP "A" and "C". The final prediction equation would, therefore, be written as follows:

$$\text{Plant height} = 61.4 + 43 (\text{linear code RFLP A}) - 7.0 (\text{quadratic code for RFLP C}).$$

Thus, if the genotypes of RFLP A and C are known for a particular plant, the height of that plant can be predicted without having to measure its height following growth to maturity. In a breeding program, the breeder can analyze the genotypes of specific RFLP of seedling plants grown in a greenhouse during the winter and need only evaluate those plants predicted to have the desired plant height in the field the following summer.

## 7.4.22.3.1 Explanation of the Mathematical Model

TABLE 6								
Hypothetical Raw data								
Genotype Observed	"A" (Linear Code)		"B" (Linear Code)		"C" (Linear Code)		(Quadratic Code)	Plant Height
1	AA	1	bb	-1	CC	1	1	100
2	aa	-1	bb	-1	cc	-1	1	0
3	Aa	0	BB	1	C	c	0	-275
4	Aa	0	Bb	0	CC	1	1	50
5	aa	-1	BB	1	CC	1	1	25

TABLE 7				
Analysis of Variance				
Source of Variance	Degrees of Freedom	RFLP locus		
		A	B	C
		Mean Squares		
Genotype	2			
Linear	1	5104	0.0	2552
Quadratic	1	44	0.0	1575
Error (deviations from regression)	2	625	6250	1458
r <sup>2</sup>		0.89	0.0	0.53
Genotypes	(code)	Means		
Homozygous Parent 1	(-1)	12.5	50	0
Heterozygote	(0)	62.5	50	75
Homozygous Parent 2	(+1)	100.0	50	58
Slope(b1)		45	0	23

#### 7.4.22.3.2 Analyzing Result Using Mathematical Model

The breeding value of an RFLP as an indirect selection criteria is a function of the additive genetic correlation between the RFLP marker and a QTL. This genetic association is presumed to be due to linkage disequilibrium rather than due to pleiotropism. The problem of recombination between an RFLP and an associated QTL can be minimized if two RFLP are identified which flank the QTL. In that instance, the probability of a double crossover would be, assuming no interference, the product of their recombination frequencies. Nevertheless, localizing a target QTL between a pair of linked RFLP can be problematic, and the complexity of the analyses increased. One possible solution to the analysis using flanking RFLPs is to use multivariate analysis and derive one or more orthogonal vectors which include information from linked (correlated) RFLP loci. Alternatively, if crossovers are detected between the two flanking RFLPs for specific entries, the RFLPs linked to the QTL can be determined and information from the other RFLP discounted.

The rate of gain from indirect selection in a population is a function of the magnitude of the phenotypic variance of the desired trait, the selection differential, the heritability of the indirect criteria, and the genetic correlation between the direct and indirect criteria. Indirect selection for RFLPs will have an advantage over direct selection if the heritability of RFLPs is higher than the desired character and the additive genetic correlation between them is high.

The "heritability" of the RFLP phenotype is 1.0, i.e., genotype=phenotype. Hence, if the correlation between RFLPs and the desired trait is greater than the heritability of the desired trait, then RFLP-facilitated selection can be advantageous. In the evaluation of 2-tridecanone ("2TD") mediated insect resistance in tomatoes, for example, the development of the colorimetric assay has contributed to increased efficiency of selection.

As noted above, quantitative differences between genotypes are usually, but not always, influenced by genes at many loci, the effects of which are small in relation to the variation arising from other causes (Falconer, D.C., "Introduction to Quantitative Genetics" (2d Edition 1981)). Consequently, the individual genes involved in the



expression of a quantitative trait are difficult to identify and Mendelian analysis cannot be applied. Id. Quantitative genetic analysis has therefore focused on estimation of breeding value which is the sum of the effects of the alleles at many loci. Nevertheless, a basic premise of quantitative genetics is that the laws which govern the inheritance of quantitative loci are the same as those which govern qualitative loci. The magnitude of the individual allelic effects which can be resolved through RFLP analysis will be a function of experimental error and recombination frequencies. As RFLP maps are developed which more completely saturate the genome, identification of more loci with smaller individual effects should be possible. Thus, RFLP analysis also offers the opportunity to determine the effects of individual loci (alleles) with major effects, and thereby to reduce the analysis of complex quantitative traits to classical Mendelian segregation ratios of individual alleles.

## 8. MUTAGENIZING ORGANISMS - ENGINEERING ASPECTS

### 8.1.1 GENERAL CONSIDERATIONS

In one aspect, this invention applies the technical field of molecular genetics to evolve the genomes of cells and organisms to acquire new and improved properties.

Cells have a number of well-established uses in molecular biology. For example, cells are commonly used as hosts for manipulating DNA in processes such as transformation and recombination. Cells are also used for expression of recombinant proteins encoded by DNA transformed into the cells. Some types of cells are also used as progenitors for generation of transgenic animals and plants. Although all of these processes are now routine, in general, the genomes of the cells used in these processes have evolved little from the genomes of natural cells, and particularly not toward acquisition of new or improved properties for use in the above processes.

The traditional approach to artificial or forced molecular evolution focuses on optimization of an individual gene having a discrete and selectable phenotype. The strategy is to clone a gene, identify a discrete function for the gene and an assay by which it can be selected, mutate selected positions in the gene (e.g., by error-prone PCR or cassette mutagenesis) and select variants of the gene for improvement in the known function of the gene. A variant having improved function can then be expressed in a desired cell type. This approach has a number of limitations. First, it is only applicable to genes that have been isolated and functionally characterized. Second, the approach is usually only applicable to genes that have a discrete function. In other words, multiple genes that cooperatively confer a single phenotype cannot usually be optimized in this manner.

Probably, most genes do have cooperative functions. Finally, this approach can only explore a very limited number of the total number of permutations even for a single gene. For example, varying even ten positions in a protein with every possible amino acid would generate  $20^{10}$  variants, which is more than can be accommodated by existing methods of transfection and screening.

In view of these limitations, the traditional approach is inadequate for improving

cellular genomes in many useful properties. For example, to improve a cell's capacity to express a recombinant protein might require modification in any or all of a substantial number of genes, known and unknown, having roles in transcription, translation, posttranslational modification, secretion or proteolytic degradation, among others. Attempting individually to optimize even all the known genes having such functions would be a virtually impossible task, let alone optimizing hitherto unknown genes which may contribute to expression in manners not yet understood.

The present invention provides inter alia novel methods for evolving the genome of whole cells and organisms which overcome the difficulties and limitations of prior methods.

This ability to evolve genes artificially is of fundamental importance. For example, cells have a number of well-established uses in molecular biology, medicine and industrial processes. For example, cells are commonly used as hosts for manipulating DNA in processes such as transformation and recombination. Cells are used for expression of recombinant proteins encoded by DNA transformed/transfected or otherwise introduced into the cells. Some types of cells are used as progenitors for generation of transgenic animals and plants. The genomes of the cells used in these processes had evolved little from the genomes of natural cells, and particularly not toward acquisition of new or improved properties for use in the above processes.

Additional methods of recursively recombining nucleic acids in vivo and selecting resulting recombinants would be of use. The present invention provides a number of new and valuable methods and compositions for whole and partial genome evolution.

Metabolic engineering is the manipulation of intermediary metabolism through the use of both classical genetics and genetic engineering techniques. Cellular engineering is generally a more inclusive term referring to the modification of cellular properties. Cameron et al. (Applied Biochem. Biotech. 38:105-140 (1993)) provide a summary of equivalent terms to describe this type of engineering, including "metabolic engineering",

which is most often used in the context of industrial microbiology and bioprocess engineering, "in vitro evolution" or "directed evolution", most often used in the context of environmental microbiology, "molecular breeding", most often used by Japanese researchers, "cellular engineering", which is used to describe modifications of bacteria, animal, and plant cells, "rational strain development", and "metabolic pathway evolution". In this application, the terms "metabolic engineering" and "cellular engineering" are used preferentially for clarity; the term "evolved" genes is used as discussed below.

Metabolic engineering can be divided into two basic categories: modification of genes endogenous to the host organism to alter metabolite flux and introduction of foreign genes into an organism. Such introduction can create new metabolic pathways leading to modified cell properties including but not limited to synthesis of known compounds not normally made by the host cell, production of novel compounds (e.g. polymers, antibiotics, etc.) and the ability to utilize new nutrient sources. Specific applications of metabolic engineering can include the production of specialty and novel chemicals, including antibiotics, extension of the range of substrates used for growth and product formation, the production of new catabolic activities in an organism for toxic chemical degradation, and modification of cell properties such as resistance to salt and other environmental factors.

Bailey (Science 252:1668-1674 (1991)) describes the application of metabolic engineering to the recruitment of heterologous genes for the improvement of a strain, with the caveat that such introduction can result in new compounds that may subsequently undergo further reactions, or that expression of a heterologous protein can result in proteolysis, improper folding, improper modification, or unsuitable intracellular location of the protein, or lack of access to required substrates. Bailey recommends careful configuration of a desired genetic change with minimal perturbation of the host. Liao (Curr. Opin. Biotech. 4:211-216 (1993)) reviews mathematical modeling and analysis of metabolic pathways, pointing out that in many cases the kinetic parameters of enzymes are unavailable or inaccurate.

Stephanopoulos et al. (Trends. Biotechnol. 11:392-396 (1993)) describe attempts to improve productivity of cellular systems or effect radical alteration of the flux through

primary metabolic pathways as having difficulty in that control architectures at key branch points have evolved to resist flux changes. They conclude that identification and characterization of these metabolic nodes is a prerequisite to rational metabolic engineering. Similarly, Stephanopoulos (Curr. Opin. Biotech. 5:196-200 (1994)) concludes that rather than modifying the "rate limiting step" in metabolic engineering, it is necessary to systematically elucidate the control architecture of bioreaction networks. The present invention is generally directed to the evolution of new metabolic pathways and the enhancement of bioprocessing through a process herein termed recursive sequence recombination. Recursive sequence recombination entails performing iterative cycles of recombination and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer, Bio/Technolog 13:549-553 (1995)). Such techniques do not require the extensive analysis and computation required by conventional methods for metabolic engineering. Recursive sequence recombination allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to traditional, pair wise recombination events.

Thus, because metabolic and cellular engineering can pose the particular problem of the interaction of many gene products and regulatory mechanisms, recursive sequence recombination (RSR) techniques provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result, whether that result is increased yield of a metabolite, altered catalytic activity or substrate specificity of an enzyme or an entire metabolic pathway, or altered response of a cell to its environment.

### 8.1.2 THE EVOLUTIONARY IMPORTANCE OF RECOMBINATION

Strain improvement is the directed evolution of an organism to be more "fit" for a desired task. In nature, adaptation is facilitated by sexual recombination. Sexual recombination allows a population to exploit the genetic diversity within it, e.g., by consolidating useful mutations and discarding deleterious ones. In this way, adaptation and evolution can proceed in leaps. In the absence of a sexual cycle, members of a population

must evolve independently by accumulating random mutations sequentially. Many useful mutations are lost while deleterious mutations can accumulate. Adaptation and evolution in this way proceeds slowly as compared to sexual evolution.

Asexual evolution is a slow and inefficient process.

Populations move as individuals rather than as groups. A diverse population is generated by the mutagenesis of a single parent resulting in a distribution of fit and unfit individuals. In the absence of a sexual cycle, each piece of genetic information of the surviving population remains in the individual mutants. Selection of the "fittest" results in many "fit" individuals being discarded along with the useful genetic information they carry. Asexual evolution proceeds one genetic event at a time and is thus limited by the intrinsic value of a single genetic event. Sexual evolution moves more quickly and efficiently. Mating within a population consolidates genetic information within the population and results in useful mutations being combined together. The combining of useful genetic information results in progeny that are much more fit than their parents. Sexual evolution thus proceeds much faster by multiple genetic events.

Years of plant and animal breeding has demonstrated the power of employing sexual recombination to effect the rapid evolution of complex genomes towards a particular task. This general principle is further demonstrated by using DNA stochastic &/or non-stochastic mutagenesis to recombine DNA molecules in vitro to accelerate the rate of directed molecular evolution. The strain improvement efforts of the fermentation industry rely on the directed evolution of microorganisms by sequential random mutagenesis. Incorporation of recombination into this iterative process greatly accelerates the strain improvement process, which in turn increases the profitability of current fermentation processes and facilitates the development of new products.

### 8.1.2.1 DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS VS NATURAL RECOMBINATION

DNA stochastic &/or non-stochastic mutagenesis includes the recursive recombination of DNA sequences. A significant difference between DNA stochastic &/or non-stochastic mutagenesis and natural sexual recombination is that DNA stochastic &/or non-stochastic mutagenesis can produce DNA sequences originating from multiple parental sequences while sexual recombination produces DNA sequences originating from only two parental sequences.

The rate of evolution is in part limited by the number of useful mutations that a member of a population can accumulate between selection events.

In sequential random mutagenesis, useful mutations are accumulated one per selection event. Many useful mutations are discarded each cycle in favor of the best performer, and neutral or deleterious mutations which survive are as difficult to lose as they were to gain and thus accumulate. In sexual evolution pairwise recombination allows mutations from two different parents to segregate and recombine in different combinations. Useful mutations can accumulate and deleterious mutations can be lost. Poolwise recombination, such as that effected by DNA stochastic &/or non-stochastic mutagenesis, has the same advantages as pairwise recombination but allows mutations from many parents to consolidate into a single progeny. Thus poolwise recombination provides a means for increasing the number of useful mutations that can accumulate each selection event. One can plot the potential number of mutations an individual can accumulate by each of these processes. Recombination is exponentially superior to sequential random mutagenesis, and this advantage increases exponentially with the number of parents that can recombine. Sexual recombination is thus more conservative. In nature, the pairwise nature of sexual recombination may provide important stability within a population by impeding the large changes in DNA sequence that can result from poolwise recombination. For the purposes of directed evolution, however, poolwise recombination is more efficient.

The potential diversity that can be generated from a population is greater as a result of poolwise recombination as compared to that resulting from pairwise recombination. Further, poolwise recombination enables the combining of multiple beneficial mutations originating from multiple parental sequences. To demonstrate the importance of poolwise recombination vs pairwise recombination in the generation of molecular diversity consider the breeding of ten independent DNA sequences each containing only one unique mutation. There are  $2^{10} = 1024$  different combinations of those ten mutations ranging from a single sequence having no mutations (the consensus) to that having all ten mutations. If this pool were recombined together by pairwise recombination, a population containing the consensus, the parents, and the 45 different combinations of any two of the mutations would result in 56 or ca. 5% of the possible 1024 mutant combinations. Alternatively, if the pool were recombined together in a poolwise fashion, all 1024 would be theoretically generated, resulting in an approximately 20 fold increase in library diversity. When looking for a unique solution to a problem in molecular evolution, the more complex the library, the more complex the possible solution. Indeed, the most fit member of a stochastic &/or non-stochastic mutagenized library often contains several mutations originating from several independent starting sequences.

#### **8.1.2.2 DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS PROVIDES RECURSIVE PAIRWISE RECOMBINATION**

In vitro DNA stochastic &/or non-stochastic mutagenesis results in the efficient production of combinatorial genetic libraries by catalyzing the recombination of multiple DNA sequences. While the result of DNA stochastic &/or non-stochastic mutagenesis is a population representing the poolwise recombination of multiple sequences, the process does not rely on the recombination of multiple DNA sequences simultaneously, but rather on their recursive pairwise recombination. The assembly of complete genes from a mixed pool of small gene fragments requires multiple annealing and elongation cycles, the thermal cycles of the primerless PCR reaction. During each thermal cycle many pairs of fragments anneal and are extended to form a combinatorial population of larger chimeric DNA fragments. After the first cycle of stochastic &/or non-stochastic mutagenesis,



chimeric fragments contain sequence originating from predominantly two different parent genes, with all possible pairs of "parental" sequence theoretically represented. This is similar to the result of a single sexual cycle within a population. During the second cycle, these chimeric fragments anneal with each other or with other small fragments, resulting in chimeras originating from up to four of the different starting sequences, again with all possible combinations of the four parental sequences theoretically represented. This second cycle is analogous to the entire population resulting from a single sexual cross, both parents and offspring, inbreeding.

Further cycles result in chimeras originating from 8, 16, 32, etc parental sequences and are analogous to further inbreedings of the preceding population. This could be considered similar to the diversity generated from a small population of birds that are isolated on an island, breeding with each other for many generations. The result mimics the outcome of "poolwise" recombination, but the path is via recursive pairwise recombination. For this reason, the DNA molecules generated from in vitro DNA stochastic &/or non-stochastic mutagenesis are not the "progeny" of the starting "parental" sequences, but rather the great, great great, great<sub>n</sub>, (n = number of thermal cycles) grand progeny of the starting "ancestor" molecules.

### 8.1.3 DEFINITIONS

The term "cognate" refers to a gene sequence that is evolutionarily and functionally related between species. For example, in the human genome, the human CD4 gene is the cognate gene to the mouse CD4 gene, since the sequences and structures of these two genes indicate that they are homologous and that both genes encode a protein which functions in signaling T-cell activation through MHC class II-restricted antigen recognition. Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker or phenotype (or a selected level of marker or phenotype), and then physically separates the cells having the desired property. Selection is a form of screening in which identification and physical separation are achieved simultaneously by expression of a selection marker, which, in some genetic

circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include luciferase, P-galactosidase, and green fluorescent protein. Selection markers include drug and toxin resistance genes.

An exogenous DNA segment is one foreign (or heterologous) to the cell or homologous to the cell but in a position within the host cell nucleic acid in which the element is not ordinarily found. Exogenous DNA segments can be expressed to yield exogenous polypeptides.

The term "**gene**" is used broadly to refer to any segment of DNA associated with a biological function. Thus, genes include coding sequences and/or the regulatory sequences required for their expression. Genes also include nonexpressed DNA segments that, for example, form recognition sequences for other proteins.

The terms "**identical**" or "**percent identity**," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection.

The phrase "**substantially identical**," in the context of two nucleic acids or polypeptides, refers to two or more sequences or subsequences that have at least 60%, preferably 80%, most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Preferably, the substantial identity exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues. In a most preferred embodiment, the sequences are substantially identical over the entire length of the coding regions.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, *Adv. Appl Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J Mol Biol* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of algorithms GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI.

Another example of a useful alignment algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments to show relationship and percent sequence identity. It also plots a tree or dendrogram showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Doolittle, *J Mol. Evol.* 35:351 - 360 (1987). The method used is similar to the method described by Higgins & Sharp, *CABIOS* 5:151-153 (1989). The program can align up to 300 sequences, each of a maximum length of 5,000 nucleotides or amino acids. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster is then aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences are aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison and by designating the program parameters. For example, a reference

sequence can be compared to other test sequences to determine the percent sequence identity relationship using the following parameters: default gap weight (3.00), default gap length weight (0.10), and weighted end gaps.

Another example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul et al., *J Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length  $W$  in the query sequence, which either match or satisfy some positive-valued threshold score  $T$  when aligned with a word of the same length in a database sequence.  $T$  is referred to as the neighborhood word score threshold (Altschul et al, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters  $M$  (reward score for a pair of matching residues; always  $> 0$ ) and  $N$  (penalty score for mismatching residues; always  $< 0$ ). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity  $X$  from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters  $W$ ,  $T$ , and  $X$  determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word length ( $W$ ) of 11, an expectation ( $E$ ) of 10,  $M=5$ ,  $N=-4$ , and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a word length ( $W$ ) of 3, an expectation ( $E$ ) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g., Karlin &

Altschul, Proc. Natl. Acad. Sci. USA 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability ( $P(N)$ ), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0. 1, more preferably less than about 0. 0 1, and most preferably less than about 0.001.

A further indication that two nucleic acid sequences or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid, as described below. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

Another indication that two nucleic acid sequences are substantially identical is that the two molecules hybridize to each other under stringent conditions.

The term "**naturally-occurring**" is used to describe an object that can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally- occurring. Generally, the term naturally-occurring refers to an object as present in a non- pathological (undiseased) individual, such as would be typical for the species.

Asexual recombination is recombination occurring without the fusion of gametes to form a zygote.

A "**mismatch repair deficient strain**" can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of *mutS*, *mutT*, *mutH*, *mutL*, *ovrD*, *dcm*, *vsr*, *umuC*, *umuD*, *sbcB*, *recJ*, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism.

## **8.2. STRATEGIES**

### **8.2.1. EVOLVING A CELL TO ACQUIRE A DESIRED FUNCTION**

#### **8.2.1.1. DESIRED FUNCTION IS SECRETION OF A PROTEIN**

Optionally, the desired function is secretion of a protein, and the plurality of cells further comprises a construct encoding the protein. The protein is optionally inactive unless secreted, and further modified cells are optionally selected for protein function.

Optionally, the protein is toxic to the plurality of cells, unless secreted. In this case, the modified or further modified cells which evolve toward acquisition of the desired function are screened by propagating the cells and recovering surviving cells.

#### **8.2.1.2. DESIRED FUNCTION IS ENHANCED RECOMBINATION**

In some methods, the desired function is enhanced recombination. In such methods, the library of fragments sometimes comprises a cluster of genes collectively conferring recombination capacity. Screening can be achieved using cells carrying a gene encoding a marker whose expression is prevented by a mutation removable by recombination. The cells are screened by their expression of the marker resulting from removal of the mutation by recombination.

#### **8.2.1.3. DESIRED FUNCTION IS IMPROVED RESISTANCE IN PLANT CELLS**

In some methods, the plurality of cells are plant cells and the desired property is improved resistance to a chemical or microbe. The modified or further modified cells (or whole plants) are exposed to the chemical or microbe and modified or further modified cells having evolved toward the acquisition of the desired function are selected by their capacity to survive the exposure.

#### **8.2.1.4. DESIRED FUNCTION IS PREDICTING EFFICACY OF A DRUG**

##### **8.2.1.4.1. A DRUG TREATING A VIRAL INFECTION**

The invention further provides methods of predicting efficacy of a drug in treating a viral infection. Such methods entail recombining a nucleic acid segment from a virus, whose infection is inhibited by a drug, with at least a second nucleic acid segment from

the virus, the second nucleic acid segment differing from the first nucleic acid segment in at least two nucleotides, to produce a library of recombinant nucleic acid segments. Host cells are then contacted with a collection of viruses having genomes including the recombinant nucleic acid segments in a media containing the drug, and progeny viruses resulting from infection of the host cells are collected.

A recombinant DNA segment from a first progeny virus recombines with at least a recombinant DNA segment from a second progeny virus to produce a further library of recombinant nucleic acid segments. Host cells are contacted with a collection of viruses having genomes including the further library or recombinant nucleic acid segments, in media containing the drug, and further progeny viruses are produced by the host cells. The recombination and selection steps are repeated, as desired, until a further progeny virus has acquired a desired degree of resistance to the drug, whereby the degree of resistance acquired and the number of repetitions needed to acquire it provide a measure of the efficacy of the drug in treating the virus. Viruses are optionally adapted to grow on particular cell lines.

#### **8.2.1.4.2. A DRUG TREATING INFECTION BY A PATHOGENIC MICROORGANISM**

The invention further provides methods of predicting efficacy of a drug in treating an infection by a pathogenic microorganism. These methods entail delivering a library of DNA fragments into a plurality of microorganism cells, at least some of which undergo recombination with segments in the genome of the cells to produce modified microorganism cells. Modified microorganisms are propagated in a media containing the drug, and surviving microorganisms are recovered. DNA from surviving microorganisms is recombined with a further library of DNA fragments at least some of which undergo recombination with cognate segments in the DNA from the surviving microorganisms to produce further modified microorganisms cells. Further modified microorganisms are propagated in media containing the drug, and further surviving microorganisms are collected. The recombination and selection steps are repeated as needed, until a further surviving microorganism has acquired a desired degree of resistance to the drug. The

degree of resistance acquired and the number of repetitions needed to acquire it provide a measure of the efficacy of the drug in killing the pathogenic microorganism.

### **8.2.1.3. METHOD**

#### **8.2.1.3.1 MODIFY OR RECOMBINE CELLS**

In one aspect, the invention provides methods of evolving a cell to acquire a desired function. Such methods entail, e.g., introducing a library of DNA fragments into a plurality of cells, whereby at least one of the fragments undergoes recombination with a segment in the genome or an episome of the cells to produce modified cells. Optionally, these modified cells are bred to increase the diversity of the resulting recombined cellular population. The modified cells, or the recombined cellular population are then screened for modified or recombined cells that have evolved toward acquisition of the desired function. DNA from the modified cells that have evolved toward the desired function is then optionally recombined with a further library of DNA fragments, at least one of which undergoes recombination with a segment in the genome or the episome of the modified cells to produce further modified cells. The further modified cells are then screened for further modified cells that have further evolved toward acquisition of the desired function. Steps of recombination and screening/selection are repeated as required until the further modified cells have acquired the desired function. In one preferred embodiment, modified cells are recursively recombined to increase diversity of the cells prior to performing any selection steps on any resulting cells.

#### **8.2.1.3.2 COAT WITH RecA**

In some methods, the library or further library of DNA fragments is coated with recA protein to stimulate recombination with the segment of the genome. The library of fragments is optionally denatured to produce single-stranded DNA, which are annealed to produce duplexes, some of which contain mismatches at points of variation in the fragments. Duplexes containing mismatches are optionally selected by affinity chromatography to immobilized MutS.



### 8.2.1.3.3 PERFORM IN VIVO RECOMBINATION

The invention further provides methods for performing in vivo recombination. At least first and second segments from at least one gene are introduced into a cell, the segments differing from each other in at least two nucleotides, whereby the segments recombine to produce a library of chimeric genes. A chimeric gene is selected from the library having acquired a desired function.

The invention further provides methods of evolving a cell to acquire a desired function. These methods entail providing a populating of different cells. The cells are cultured under conditions whereby DNA is exchanged between cells, forming cells with hybrid genomes. The cells are then screened or selected for cells that have evolved toward acquisition of a desired property. The DNA exchange and screening/selecting steps are repeated, as needed, with the screened/selected cells from one cycle forming the population of different cells in the next cycle, until a cell has acquired the desired property.

Mechanisms of DNA exchange include conjugation, phage-mediated transduction, liposome delivery, protoplast fusion, and sexual recombination of the cells. Optionally, a library of DNA fragments can be transformed or electroporated into the cells.

### 8.2.1.3.4 PROTOPLAST-MEDIATED EXCHANGE

As noted, some methods of evolving a cell to acquire a desired property are effected by protoplast-mediated exchange of DNA between cells. Such methods entail forming protoplasts of a population of different cells. The protoplasts are then fused to form hybrid protoplasts, in which genomes from the protoplasts recombine to form hybrid genomes. The hybrid protoplasts are incubated under conditions promoting regeneration of cells. The regenerated cells can be recombined one or more times (i.e., via protoplasting or any other method than combines genomes of cells) to increase the diversity of any resulting cells. Preferably, regenerated cells are recombined several times, e.g., by protoplast fusion to generate a diverse population of cells. The next step is to select or screen to isolate regenerated cells that have evolved toward acquisition of the desired

property. DNA exchange and selection/screening steps are repeated, as needed, with regenerated cells in one cycle being used to form protoplasts in the next cycle until the regenerated cells have acquired the desired property. Industrial microorganisms are a preferred class of organisms for conducting the above methods. Some methods further comprise a step of selecting or screening for fused protoplasts free from unfused protoplasts of parental cells. Some methods further comprise a step of selecting or screening for fused protoplasts with hybrid genomes free from cells with parental genomes. In some methods, protoplasts are provided by treating individual cells, mycelia or spores with an enzyme that degrades cell walls. In some methods, the strain is a mutant that is lacking capacity for intact cell wall synthesis, and protoplasts form spontaneously. In some methods, protoplasts are formed by treating growing cells with an inhibitor of cell wall formation to generate protoplasts. In some methods, the desired property is expression and/or secretion of a protein or secondary metabolite, such as an industrial enzyme, a therapeutic protein, a primary metabolite such as lactic acid or ethanol, or a secondary metabolite such as erythromycin cyclosporin A or taxol. In other methods it is the ability of the cell to convert compounds provided to the cell to different compounds. In yet other methods, the desired property is capacity for meiosis. In some methods, the desired property is compatibility to form a heterokaryon with another strain.

The invention further provides methods of evolving a cell toward acquisition of a desired property. These methods entail providing a population of different cells. DNA is isolated from a first subpopulation of the different cells and encapsulated in liposomes. Protoplasts are formed from a second subpopulation of the different cells. Liposomes are fused with the protoplasts, whereby DNA from the liposomes is taken up by the protoplasts and recombines with the genomes of the protoplasts. The protoplasts are incubated under regenerating conditions. Regenerating or regenerated cells are then selected or screened for evolution toward the desired property.

#### **8.2.1.3.4 REITERATIVE POOLING AND BREEDING OF HIGHER ORGANISMS**

The method also provides methods of reiterative pooling and breeding of higher organisms. In the methods, a library of diverse multicellular organisms are produced (e.g.,

plants, animals or the like). A pool of male gametes is provided along with a pool of female gametes. At least one of the male pool or the female pool comprises a plurality of different gametes derived from different strains of a species or different species. The male gametes are used to fertilize the female gametes. At least a portion of the resulting fertilized gametes grow into reproductively viable organisms. These reproductively viable organisms are crossed (e.g., by pairwise pooling and joining of the male and female gametes as before) to produce a library of diverse organisms. The library is then selected for a desired trait or property.

The library of diverse organisms can comprise a plurality of plants such as Gramineae, Fetucoeidae, Poacoideae, Agrostis, Phleum, Dactylis, Sorgum, Setaria, Zea, Oryza, Triticum, Secale, Avena, Hordeum, Saccharum, Poa, Festuca, Stenotaphrum, Cynodon, Coix, Olyreae, Phareae, Compositae or Leguminosae. For example, the plants can be e.g., corn, rice, wheat, rye, oats, barley, pea, beans, lentil, peanut, yam bean, cowpeas, velvet beans, soybean, clover, alfalfa, lupine, vetch, lotus, sweet clover, wisteria, sweet pea, sorghum, millet, sunflower, canola or the like.

Similarly, the library of diverse organisms can include a plurality of animals such as non-human mammals, fish, insects, or the like.

Optionally, a plurality of selected library members can be crossed by pooling gametes from the selected members and repeatedly crossing any resulting additional reproductively viable organisms to produce a second library of diverse organisms (e.g., by split pair wise pooling and rejoining of the male and female gametes). Here again, the second library can be selected for a desired trait or property, with the resulting selected members forming the basis for additional poolwise breeding and selection. A feature of the invention is the libraries made by these (or any preceding) method.

### **8.3. ORIGIN OF CELLS**

#### **8.3.1 EMBRYONIC CELLS OF AN ANIMAL**

In some methods, the plurality of cells are embryonic cells of an animal, and the method further comprises propagating the transformed cells to transgenic animals. The

plurality of cells can be a plurality of industrial microorganisms that are enriched for microorganisms which are tolerant to desired process conditions (heat, light, radiation, selected pFL presence of detergents or other denaturants, presence of alcohols or other organic molecules, etc.).

### 8.3.2 ARTIFICIAL CHROMOSOMES

The invention further provides methods of evolving a cell toward acquisition of a desired property using artificial chromosomes. Such methods entail introducing a DNA fragment library cloned into an artificial chromosome into a population of cells. The cells are then cultured under conditions whereby sexual recombination occurs between the cells, and DNA fragments cloned into the artificial chromosome recombines by homologous recombination with corresponding segments of endogenous chromosomes of the populations of cells, and endogenous chromosomes recombine with each other. Cells can also be recombined via conjugation. Any resulting cells can be recombined via any method noted herein, as many times as desired, to generate a desired level of diversity in the resulting recombinant cells. In any case, after generating a diverse library of cells, the cells that have evolved toward acquisition of the desired property are screened and/or selected for a desired property. The method is then repeated with cells that have evolved toward the desired property in one cycle forming the population of different cells in the next cycle. Here again, multiple cycles of in vivo recombination are optionally performed prior to any additional selection or screening steps.

The invention further provides methods of evolving a DNA segment cloned into an artificial chromosome for acquisition of a desired property. These methods entail providing a library of variants of the segment, each variant cloned into separate copies of an artificial chromosome. The copies of the artificial chromosome are introduced into a population of cells. The cells are cultured under conditions whereby sexual recombination occurs between cells and homologous recombination occurs between copies of the artificial chromosome bearing the variants. Variants are then screened or selected for evolution toward acquisition of the desired property.

The invention further provides hyperrecombinogenic recA proteins.

#### **8.4. METHOD TO ACQUIRE A BIOCATALYTIC ACTIVITY**

One aspect of the invention is a method of evolving a biocatalytic activity of a cell, comprising:

(a) recombining at least a first and second DNA segment from at least one gene conferring ability to catalyze a reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(b) screening at least one recombinant gene from the library that confers enhanced ability to catalyze the reaction of interest by the cell relative to a wild type form of the gene;

(c) recombining at least a segment from at least one recombinant gene with a further DNA segment from at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(d) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze the reaction of interest in the cell relative to a previous recombinant gene;

(e) repeating (c) and (d), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze the reaction of interest by the cell.

##### **8.4.1. METHOD TO EVOLVE A GENE TO CATALYZE A RXN OF INTEREST**

Another aspect of the invention is a method of evolving a gene to confer ability to catalyze a reaction of interest, the method comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to catalyze a reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers enhanced

ability to catalyze a reaction of interest relative to a wild type form of the gene;

(3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze a reaction of interest relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze a reaction of interest.

#### **8.4.2. METHOD TO GENERATE A NEW BIOCATALYTIC ACTIVITY IN A CELL**

A further aspect of the invention is a method of generating a new biocatalytic activity in a cell, comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to catalyze a first reaction related to a second reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers a new ability to catalyze the second reaction of interest;

(3) recombining at least a segment from at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze the second reaction of interest in the cell relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze the second reaction of interest in the cell.

#### **8.4.3. METHOD TO MODIFY A METABOLIC PATHWAY EVOLVED BY RECURSIVE SEQUENCE RECOMBINATION**

Another aspect of the invention is a modified form of a cell, wherein the modification comprises a metabolic pathway evolved by recursive sequence recombination.

A further aspect of the invention is a method of optimizing expression of a gene product, the method comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to produce the gene product, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers optimized expression of the gene product relative to a wild type form of the gene;

(3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers optimized ability to produce the gene product relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of optimized ability to express the gene product.

#### **8.4.4. METHOD TO EVOLVE A BIOSENSOR**

A further aspect of the invention is a method of evolving a biosensor for a compound A of interest, the method comprising:

- (1) recombining at least first and second DNA segments from at least one gene conferring ability to detect a related compound B, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;
- (2) screening at least one recombinant gene from the library that confers optimized ability to detect compound A relative to a wild type form of the gene;
- (3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;
- (4) screening at least one further recombinant gene from the further library of recombinant genes that confers optimized ability to detect compound A relative to a previous recombinant gene;
- (5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of optimized ability to detect compound A.

## 8.5. FERMENTATION OF MICRO-ORGANISMS

The fermentation of microorganisms for the production of natural products is the oldest and most sophisticated application of biocatalysis. Industrial microorganisms effect the multistep conversion of renewable feedstocks to high value chemical products in a single reactor and in so doing catalyze a multi-billion dollar industry. Fermentation products range from fine and commodity chemicals such as ethanol, lactic acid, amino acids and vitamins, to high value small molecule pharmaceuticals, protein pharmaceuticals, and industrial enzymes. (See, e.g., McCoy (1998) C&EN 13-19) for an introduction to biocatalysis. Success in bringing these products to market and success in competing in the market depends on continuous improvement of the whole cell biocatalysts. Improvements include increased yield of desired products, removal of unwanted co-metabolites, improved utilization of inexpensive carbon and nitrogen sources, and adaptation to fermenter conditions, increased production of a primary metabolite, increased production of a secondary metabolite, increased tolerance to acidic



conditions, increased tolerance to basic conditions, increased tolerance to organic solvents, increased tolerance to high salt conditions and increased tolerance to high or low temperatures. Shortcomings in any of these areas can result in high manufacturing costs, inability to capture or maintain market share, and failure of bringing promising products to market. For this reason, the fermentation industry invests significant financial and personnel resources in the improvement of production strains. Current strategies for strain improvement rely on the empirical and iterative modification of fermenter conditions and genetic manipulation of the producing organism.

While advances in the molecular biology of established industrial organisms have been made, rational metabolic engineering is information intensive and is not broadly applicable to less characterized industrial strains. The most widely practiced strategy for strain improvement employs random mutagenesis of the producing strain and screening for mutants having improved properties. For mature strains, those subjected to many rounds of improvement, these efforts routinely provide a 10% increase in product titre per year. Although effective, this classic strategy is slow, laborious, and expensive. Technological advances in this area are aimed at automation and increasing sample screening throughput in hopes of reducing the cost of strain improvement. However, the real technical barrier resides in the intrinsic limitation of single mutations to effect significant strain improvement. The methods herein overcome this limitation and provide access to multiple useful mutations per cycle which can be used to complement automation technologies and catalyze strain improvement processes.

The methods herein allow biocatalysts to be improved at a faster pace than conventional methods. Whole genome stochastic &/or non-stochastic mutagenesis can at least double the rate of strain improvement for microorganisms used in fermentation as compared to traditional methods. This provides for a relative decrease in the cost of fermentation processes. New products can enter the market sooner, producers can increase profits as well as market share, and consumers gain access to more products of higher quality and at lower prices. Further, increased efficiency of production processes translates to less waste production and more frugal use of resources. Whole genome stochastic &/or

non-stochastic mutagenesis provides a means of accumulating multiple useful mutation per cycle and thus eliminate the inherent limitation of current strain improvement programs (SIPs).

DNA stochastic &/or non-stochastic mutagenesis provides recursive mutagenesis, recombination, and selection of DNA sequences. A key difference between DNA stochastic &/or non-stochastic mutagenesis-mediated recombination and natural sexual recombination is that DNA stochastic &/or non-stochastic mutagenesis effects both the pairwise (two parents) and the poolwise (multiple parents) recombination of parent molecules, as described Supra. Natural recombination is more conservative and is limited to pairwise recombination. In nature, pairwise recombination provides stability within a population by preventing large leaps in sequences or genomic structure that can result from poolwise recombination. However, for the purposes of directed evolution, poolwise recombination is appealing since the beneficial mutations of multiple parents can be combined during a single cross to produce a superior offspring. Poolwise recombination is analogous to the crossbreeding of inbred strains in classic strain improvement, except that the crosses occur between many strains at once. In essence, poolwise recombination is a sequence of events that effects the recombination of a population of nucleic acid sequences that results in the generation of new nucleic acids that contains genetic information from more than two of the original nucleic acids. The power of in vitro DNA stochastic &/or non-stochastic mutagenesis is that large combinatorial libraries can be generated from a small pool of DNA fragments stochastic &/or non-stochastic mutagenized by recursive pairwise annealing and extension reactions, "matings." Many of the in vivo recombination formats described (such as plasmid-plasmid, plasmid-chromosome, phage-phage, phage-chromosome, phage-plasmid, conjugal DNA-chromosome, exogenous DNA-chromosome, chromosome-chromosome, with the DNA being introduced into the cell by natural and non-natural competence, transduction, transfection, conjugation, protoplast fusion, etc.) result primarily in the pairwise recombination of two DNA molecules. Thus, these formats when executed for only a single cycle of recombination are inherently limited in their potential to generate molecular diversity. To generate the level of diversity obtained by in vitro DNA stochastic

&/or non-stochastic mutagenesis methods, pairwise mating formats must be carried out recursively, i.e for many generations, prior to screening for improved sequences.

Thus a pool of DNA sequences, such as four independent chromosomes, must be recombined, for example by protoplast fusion, and the progeny of that recombination (each representing a unique outcome of the pairwise mating) must then be pooled, without selection, and then recombined again, and again, and again. This process should be repeated for a sufficient number of cycles to result in progeny having the desired complexity. Only once sufficient diversity has been generated, should the resulting population be screened for new and improved sequences.

There are a few general methods for effecting efficient recombination in prokaryotes. Bacteria have no known sexual cycle per se, but there are natural mechanisms by which the genomes of these organisms undergo recombination. These mechanisms include natural competence, phage-mediated transduction, and cell-cell conjugation. Bacteria that are naturally competent are capable of efficiently taking up naked DNA from the environment. If homologous, this DNA undergoes recombination with the genome of the cell, resulting in genetic exchange. *Bacillus subtilis*, the primary production organism of the enzyme industry, is known for the efficiency with which it carries out this process.

In generalized transduction, a bacteriophage mediates genetic exchange. A transducing phage will often package headfulls of the host genome. These phage can infect a new host and deliver a fragment of the former host genome which is frequently integrated via homologous recombination. Cells can also transfer DNA between themselves by conjugation. Cells containing the appropriate mating factors transfer episomes as well as entire chromosomes to an appropriate acceptor cell where it can recombine with the acceptor genome. Conjugation resembles sexual recombination for microbes and can be intraspecific, interspecific, and intergeneric. For example, an efficient means of transforming *Streptomyces* sp., a genera responsible for producing many commercial antibiotics, is by the conjugal transfer of plasmids from *Escherichia coli*.

For many industrial microorganisms, knowledge of competence, transducing phage, or fertility factors is lacking. Protoplast fusion has been developed as a versatile and general alternative to these natural methods of recombination. Protoplasts are prepared by removing the cell wall by treating cells with lytic enzymes in the presence of osmotic stabilizers. In the presence of a fusogenic agent, such as polyethylene glycol (PEG), protoplasts are induced to fuse and form transient hybrids or "fusants." During this hybrid state, genetic recombination occurs at high frequency allowing the genomes to reassort. The final crucial step is the successful segregation and regeneration of viable cells from the fused protoplasts. Protoplast fusion can be intraspecific, interspecific, and intergeneric and has been applied to both prokaryotes and eukaryotes. In addition, it is possible to fuse more than two cells, thus providing a mechanism for effecting poolwise recombination. While no fertility factors, transducing phages or competency development is needed for protoplast fusion, a method for the formation, fusing, and regeneration of protoplasts is typically optimized for each organism. Protoplast fusion as applied to poolwise recombination is described in more detail, *supra*.

One key to SIP is having an assay that can be dependably used to identify a few mutants out of thousands that have subtle increases in product yield. The limiting factor in many assay formats is the uniformity of cell growth. This variation is the source of baseline variability in subsequent assays. Inoculum size and culture environment (temperature/humidity) are sources of cell growth variation. Automation of all aspects of establishing initial cultures and state-of-the-art temperature and humidity controlled incubators are useful in reducing variability.

Mutant cells or spores are separated on solid media to produce individual sporulating colonies. Using an automated colony picker (Q-bot, Genefix, U.K.), colonies are identified, picked, and 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm. glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells (or mycelia) and spores. The time the pin is in the colony, the number of dips to inoculate the culture medium, and the time the pin is in that medium each effect inoculum

size, and each can be controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. The glass balls act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter.

**1. Prescreen** The ability to detect a subtle increase in the performance of a mutant over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size a prescreen that increases the number of mutants processed by 10-fold can be used. The goal of the primary screen will be to quickly identify mutants having equal or better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture. The primary screen is an agar plate screen is analyzed by the Q-bot colony picker. Although assays can be fundamentally different, many result, e.g. , in the production of colony halos. For example, antibiotic production is assayed on plates using an overlay of a sensitive indicator strain, such as *B. subtilis*. Antibiotic production is typically assayed as a zone of clearing (inhibited growth of the indicator organism) around the producing organism. Similarly, enzyme production can be assayed on plates containing the enzyme substrate, with activity being detected as a zone of substrate modification around the producing colony. Product titre is correlated with the ratio of halo area to colony area.

The Q-bot or other automated system is instructed to only pick colonies having a halo ratio in the top 10% of the population i.e. 10,000 mutants from the 100,000 entering the plate prescreen. This increases the number of improved clones in the secondary assay and eliminates the wasted effort of screening knock-out and low producers. This improves the "hit rate" of the secondary assay.

## **8.6. EXPERIMENTAL APPLICATIONS**

### **8.6.1 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

#### **8.6.1.1 GENERAL TECHNIQUES**

##### **8.6.1.1.1 STARTING MATERIALS**

Thus, a general method for recursive sequence recombination for the embodiments herein is to begin with a gene encoding an enzyme or enzyme subunit and to evolve that gene either for ability to act on a new substrate, or for enhanced catalytic properties with an old substrate, either alone or in combination with other genes in a multistep pathway. The term "gene" is used herein broadly to refer to any segment or sequence of DNA associated with a biological function. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters. The ability to use a new substrate can be assayed in some instances by the ability to grow on a substrate as a nutrient source. In other circumstances such ability can be assayed by decreased toxicity of a substrate for a host cell, hence allowing the host to grow in the presence of that substrate. Biosynthesis of new compounds, such as antibiotics, can be assayed similarly by growth of an indicator organism in the presence of the host expressing the evolved genes. For example, when an indicator organism used in an overlay of the host expressing the evolved gene(s), wherein the indicator organism is sensitive or expected to be sensitive to the desired antibiotic, growth of the indicator organism would be inhibited in a zone around the host cell or colony expressing the evolved gene(s).

Another method of identifying new compounds is the use of standard analytical techniques such as mass spectroscopy, nuclear magnetic resonance, high performance liquid chromatography, etc. Recombinant microorganisms can be pooled and extracts or media supernatants assayed from these pools. Any positive pool can then be subdivided and the procedure repeated until the single positive is identified ("sib-selection").

In some instances, the starting material for recursive sequence recombination is a discrete gene, cluster of genes, or family of genes known or thought to be associated with

metabolism of a particular class of substrates. One of the advantages of the instant invention is that structural information is not required to estimate which parts of a sequence should be mutated to produce a functional hybrid enzyme.

In some embodiments of the invention, an initial screening of enzyme activities in a particular assay can be useful in identifying candidate enzymes as starting materials. For example, high throughput screening can be used to screen enzymes for dioxygenase-type activities using aromatic acids as substrates. Dioxygenases typically transform indole-2-carboxylate and indole-3-carboxylate to colored products, including indigo (Eaton et. al. J. Bacteriol. 177:6983-6988 (1995)). DNA encoding enzymes that give some activity in the initial assay can then be recombined by the recursive techniques of the invention and rescreened. The use of such initial screening for candidate enzymes against a desired target molecule or analog of the target molecule can be especially useful to generate enzymes that catalyze reactions of interest such as catabolism of man-made pollutants.

This type of high throughput screening can also be used during each round of recursive sequence recombination to identify mutants that possess the highest level of the desired activity. For example, penicillin G acylases have been isolated by looking for clones that allow a leucine auxotroph to hydrolyse penicillin G analogue phenylacetyl-L-leucine, thereby producing leucine and allowing cell growth (Martin, L. et al., FEMS Microbiology Lett. 125:287-292 (1995)). Positives from this selection are then screened by a more labour-intensive method for ability to hydrolyse penicillin G.

This same selection on phenylacetyl-L-leucine can be used when evolving penicillin G acylase for greater activity by recursive sequence recombination. After each round of recombination the library of acylase genes is transformed into a leucine auxotroph. Those that grow fastest are picked as probably having the most active acylase. The acylases are then be tested against the real substrate, penicillin G, by a more laborious screen such as HPLC. Thus, even if there is no convenient high throughput screen for an enzyme or a metabolic pathway, it is often possible to find a rapid detection method that can approximately measure the desired phenotype, thereby reducing the numbers of colonies that must be screened more accurately.

The starting material can also be a segment of such a gene or cluster that is recombined in isolation of its surrounding DNA, but is relinked to its surrounding DNA before screening/selection of recombination products. In other instances, the starting material for recombination is a larger segment of DNA that includes a coding sequence or other locus associated with metabolism of a particular substrate at an unknown location. For example, the starting material can be a chromosome, episome, YAC, cosmid, or phage P1 clone. In still other instances, the starting material is the whole genome of an organism that is known to have desirable metabolic properties, but for which no information localizing the genes associated with these characteristics is available.

In general any type of cells can be used as a recipient of evolved genes. Cells of particular interest include many bacterial cell types, both gram-negative and gram-positive, such as *Rhodococcus*, *Streptomyces*, *Actinomyces*, *Corynebacteria*, *Penicillium*, *Bacillus*, *Escherichia coli*, *Pseudomonas*, *Salmonella*, and *Erwinia*. Cells of interest also include eukaryotic cells, particularly mammalian cells (e.g., mouse, hamster, primate, human), both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIHM), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean, sugarcane, tobacco, and arabidopsis; fish, algae, fungi (*Penicillium*, *Fusarium*, *Aspergillus*, *Podospora*, *Neurospora*), insects, yeasts (*Pichia* and *Saccharomyces*).

The choice of host will depend on a number of factors, depending on the intended use of the engineered host, including pathogenicity, substrate range, environmental hardiness, presence of key intermediates, ease of genetic manipulation, and likelihood of promiscuous transfer of genetic information to other organisms. Particularly advantageous hosts are *E. coli*, lactobacilli, *Streptomyces*, *Actinomyces* and filamentous fungi.

The breeding procedure starts with at least two substrates, which generally show substantial sequence identity to each other (i.e., at least about 50%, 70%, 80% or 90% sequence identity) but differ from each other at certain positions. The difference can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in perhaps 5-20 positions. For recombination to generate



increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second as a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. The segments can also be from nonallelic genes showing some degree of structural and is usually functional relatedness (e.g., different genes within a superfamily such as the immunoglobulin superfamily). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example, one segment can be a subsequence of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form. The starting DNA segments are recombined by any of the recursive sequence recombination formats described above to generate a diverse library of recombinant DNA segments.

Such a library can vary widely in size from having fewer than to more than  $10^5$ ,  $10^7$ , or  $10^9$  members. In general, the starting segments and the recombinant libraries generated include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. However, if this is not the case, the recombinant DNA segments in the library can be inserted into a common vector providing the missing sequences before performing screening/selection.

If the recursive sequence recombination format employed is an *in vivo* format, the library of recombinant DNA segments generated already exists in a cell, which is usually the cell type in which expression of the enzyme with altered substrate specificity is desired. If recursive sequence recombination is performed *in vitro*, the recombinant library is preferably introduced into the desired cell type before screening/selection. The members

of the recombinant library can be linked to an episome or virus before introduction or can be introduced directly. In some embodiments of the invention, the library is amplified in a first host, and is then recovered from that host and introduced to a second host more amenable to expression, selection, or screening, or any other desirable parameter. The manner in which the library is introduced into the cell type depends on the DNA-uptake characteristics of the cell type, e.g., having viral receptors, being capable of conjugation, or being naturally competent. If the cell type is insusceptible to natural and chemical-induced competence, but susceptible to electroporation, one would usually employ electroporation. If the cell type is insusceptible to electroporation as well, one can employ biolistics. The biolistic PDS-1000 Gene Gun (Biorad, Hercules, CA) uses helium pressure to accelerate DNA-coated gold or tungsten microcarriers toward target cells.

The process is applicable to a wide range of tissues, including plants, bacteria, fungi, algae, intact animal tissues, tissue culture cells, and animal embryos. One can employ electronic pulse delivery, which is essentially a mild electroporation format for live tissues in animals and patients. Zhao, *Advanced Drug Delivery Reviews* 17:257-262 (1995). After introduction of the library of recombinant DNA genes, the cells are optionally propagated to allow expression of genes to occur.

#### 8.6.1.1.2 SELECTION AND SCREENING

Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker and then physically separates the cells having the desired property. Selection is a form of screening in which identification and physical separation are achieved simultaneously, for example, by expression of a selectable marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include, for example, luciferase,  $\beta$ -galactosidase, and green fluorescent protein.

Screening can also be done by observing such aspects of growth as colony size, halo formation, etc. Additionally, screening for production of a desired compound, such as a therapeutic drug or "designer chemical" can be accomplished by observing binding of

cell products to a receptor or ligand, such as on a solid support or on a column. Such screening can additionally be accomplished by binding to antibodies, as in an ELISA. In some instances the screening process is preferably automated so as to allow screening of suitable numbers of colonies or cells. Some examples of automated screening devices include fluorescence activated cell sorting, especially in conjunction with cells immobilized in agarose (see Powell et. al. *Bio/Technology* 8:333-337 (1990); Weaver et. al. *Methods* 2:234- 247 (1991)), automated ELISA assays, scintillation proximity assays (Hart, H.E. et al., *Molecular Immunol.* 16:265-267 (1979)) and the formation of fluorescent, coloured or UV absorbing compounds on agar plates or in microtitre wells (Krawiec, S., *Devel. Indust. Microbiology* 31:103-114 (1990)).

Selectable markers can include, for example, drug, toxin resistance, or nutrient synthesis genes. Selection is also done by such techniques as growth on a toxic substrate to select for hosts having the ability to detoxify a substrate, growth on a new nutrient source to select for hosts having the ability to utilize that nutrient source, competitive growth in culture based on ability to utilize a nutrient source, etc.

In particular, uncloned but differentially expressed proteins (e.g., those induced in response to new compounds, such as biodegradable pollutants in the medium) can be screened by differential display (Appleyard et al. *Mol. Gen. Gent.* 247:338-342 (1995)). Hopwood (*Phil Trans R. Soc. Lond B* 324:549-562) provides a review of screens for antibiotic production. Omura (*Microbio. Rev.* 50:259-279 (1986) and Nisbet (*Ann Rev. Med. Chem.* 21:149-157 (1986)) disclose screens for antimicrobial agents, including supersensitive bacteria, detection of beta-lactamase and D,D- carboxypeptidase inhibition, beta-lactamase induction, chromogenic substrates and monoclonal antibody screens.

Antibiotic targets can also be used as screening targets in high throughput screening. Antifungals are typically screened by inhibition of fungal growth. Pharmacological agents can be identified as enzyme inhibitors using plates containing the enzyme and a chromogenic substrate, or by automated receptor assays. Hydrolytic enzymes (e.g., proteases, amylases) can be screened by including the substrate in an agar plate and scoring for a hydrolytic clear zone or by using a colorimetric indicator (Steele et al. *Ann. Rev. Microbiol.* 45:89-106 (1991)). This can be coupled with the use of stains to

detect the effects of enzyme action (such as congo red to detect the extent of degradation of celluloses and hemicelluloses).

Tagged substrates can also be used. For example, lipases and esterases can be screened using different lengths of fatty acids linked to umbelliferyl. The action of lipases or esterases removes this tag from the fatty acid, resulting in a quenching or enhancement of umbelliferyl fluorescence. These enzymes can be screened in microtiter plates by a robotic device.

#### 8.6.1.1.3 FACS

Fluorescence activated cell sorting (FACS) methods are also a powerful tool for selection/screening. In some instances a fluorescent molecule is made within a cell (e.g., green fluorescent protein). The cells producing the protein can simply be sorted by FACS. Gel microdrop technology allows screening of cells encapsulated in agarose microdrops (Weaver et al. *Methods* 2:234-247 (1991)). In this technique products secreted by the cell (such as antibodies or antigens) are immobilized with the cell that generated them. Sorting and collection of the drops containing the desired product thus also collects the cells that made the product, and provides a ready source for the cloning of the genes encoding the desired functions. Desired products can be detected by incubating the encapsulated cells with fluorescent antibodies (Powell et al. *Bio/Technology* 8:333-337 (1990)). FACS sorting can also be used by this technique to assay resistance to toxic compounds and antibiotics by selecting droplets that contain multiple cells (i.e., the product of continued division in the presence of a cytotoxic compound; Goguen et al. *Nature* 363:189-190 (1995)). This method can select for any enzyme that can change the fluorescence of a substrate that can be immobilized in the agarose droplet.

#### **8.6.1.1.4 REPORTER MOLECULE**

In some embodiments of the invention, screening can be accomplished by assaying reactivity with a reporter molecule reactive with a desired feature of, for example, a gene product. Thus, specific functionalities such as antigenic domains can be screened with antibodies specific for those determinants.

#### **8.6.1.1.5 CELL-CELL INDICATOR**

In other embodiments of the invention, screening is preferably done with a cell-cell indicator assay. In this assay format, separate library cells (Cell A, the cell being assayed) and reporter cells (Cell B, the assay cell) are used.

Only one component of the system, the library cells, is allowed to evolve. The screening is generally carried out in a two-dimensional immobilized format, such as on plates. The products of the metabolic pathways encoded by these genes (in this case, usually secondary metabolites such as antibiotics, polyketides, carotenoids, etc.) diffuse out of the library cell to the reporter cell. The product of the library cell may affect the reporter cell in one of a number of ways.

The assay system (indicator cell) can have a simple readout (e.g., green fluorescent protein, luciferase,  $\beta$ -galactosidase) which is induced by the library cell product but which does not affect the library cell. In these examples the desired product can be detected by colorimetric changes in the reporter cells adjacent to the library cell.

#### **8.6.1.1.6 FEEDBACK MECHANISM**

In other embodiments, indicator cells can in turn produce something that modifies the growth rate of the library cells via a feedback mechanism. Growth rate feedback can detect and accumulate very small differences. For example, if the library and reporter cells

are competing for nutrients, library cells producing compounds to inhibit the growth of the reporter cells will have more available nutrients, and thus will have more opportunity for growth. This is a useful screen for antibiotics or a library of polyketide synthesis gene clusters where each of the library cells is expressing and exporting a different polyketide gene product.

#### 8.6.1.1.7 SECRETION

Another variation of this theme is that the reporter cell for an antibiotic selection can itself secrete a toxin or antibiotic that inhibits growth of the library cell. Production by the library cell of an antibiotic that is able to suppress growth of the reporter cell will thus allow uninhibited growth of the library cell.

Conversely, if the library is being screened for production of a compound that stimulates the growth of the reporter cell (for example, in improving chemical syntheses, the library cell may supply nutrients such as amino acids to an auxotrophic reporter, or growth factors to a growth-factor- dependent reporter. The reporter cell in turn should produce a compound that stimulates the growth of the library cell. Interleukins, growth factors, and nutrients are possibilities. Further possibilities include competition based on ability to kill surrounding cells, positive feedback loops in which the desired product made by the evolved cell stimulates the indicator cell to produce a positive growth factor for cell A, thus indirectly selecting for increased product formation.

In some embodiments of the invention it can be advantageous to use a different organism (or genetic background) for screening than the one that will be used in the final product. For example, markers can be added to DNA constructs used for recursive sequence recombination to make the microorganism dependent on the constructs during the improvement process, even though those markers may be undesirable in the final recombinant microorganism.

Likewise, in some embodiments it is advantageous to use a different substrate for screening an evolved enzyme than the one that will be used in the final product. For

example, Evnin et al. (Proc. Natl. Acad. Sci. U.S.A. 87:6659-6663 (1990)) selected trypsin variants with altered substrate specificity by requiring that variant trypsin generate an essential amino acid for an arginine auxotroph by cleaving arginine -naphthylamide. This is thus a selection for arginine-specific trypsin, with the growth rate of the host being proportional to that of the enzyme activity.

The pool of cells surviving screening and/or selection is enriched for recombinant genes conferring the desired phenotype (e.g. altered substrate specificity, altered biosynthetic ability, etc.). Further enrichment can be obtained, if desired, by performing a second round of screening and/or selection without generating additional diversity.

The recombinant gene or pool of such genes surviving one round of screening/selection forms one or more of the substrates for a second round of recombination. Again, recombination can be performed in vivo or in vitro by any of the recursive sequence recombination formats described above.

If recursive sequence recombination is performed in vitro, the recombinant gene or genes to form the substrate for recombination should be extracted from the cells in which screening/selection was performed. Optionally, a subsequence of such gene or genes can be excised for more targeted subsequent recombination. If the recombinant gene(s) are contained within episomes, their isolation presents no difficulties. If the recombinant genes are chromosomally integrated, they can be isolated by amplification primed from known sequences flanking the regions in which recombination has occurred. Alternatively, whole genomic DNA can be isolated, optionally amplified, and used as the substrate for recombination. Small samples of genomic DNA can be amplified by whole genome amplification with degenerate primers (Barrett et al. Nucleic Acids Research 23:3488-3492 (1995)). These primers result in a large amount of random 3' ends, which can undergo homologous recombination when reintroduced into cells.

If the second round of recombination is to be performed in vivo, as is often the case, it can be performed in the cell surviving screening/selection, or the recombinant genes can be transferred to another cell type (e.g., a cell is type having a high frequency of

mutation and/or recombination). In this situation, recombination can be effected by introducing additional DNA segment(s) into cells bearing the recombinant genes. In other methods, the cells can be induced to exchange genetic information with each other by, for example, electroporation. In some methods, the second round of recombination is performed by dividing a pool of cells surviving screening/selection in the first round into two subpopulations. DNA from one subpopulation is isolated and transfected into the other population, where the recombinant gene(s) from the two subpopulations recombine to form a further library of recombinant genes. In these methods, it is not necessary to isolate particular genes from the first subpopulation or to take steps to avoid random shearing of DNA during extraction. Rather, the whole genome of DNA sheared or otherwise cleaved into manageable sized fragments is transfected into the second subpopulation. This approach is particularly useful when several genes are being evolved simultaneously and/or the location and identity of such genes within chromosome are not known.

The second round of recombination is sometimes performed exclusively among the recombinant molecules surviving selection. However, in other embodiments, additional substrates can be introduced. The additional substrates can be of the same form as the substrates used in the first round of recombination, i.e., additional natural or induced mutants of the gene or cluster of genes, forming the substrates for the first round. Alternatively, the additional substrate(s) in the second round of recombination can be exactly the same as the substrate(s) in the first round of replication.

After the second round of recombination, recombinant genes conferring the desired phenotype are again selected. The selection process proceeds essentially as before. If a suicide vector bearing a selective marker was used in the first round of selection, the same vector can be used again. Again, a cell or pool of cells surviving selection is selected. If a pool of cells, the cells can be subject to further enrichment.



## 8.6.1.2 GENERAL METHODS

### 8.6.1.2.1 IN VITRO

In Vitro Formats one format for recursive sequence recombination in vitro is illustrated herein. The initial substrates for recombination are a pool of related sequences. The X's show where the sequences diverge. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or stochastic &/or non-stochastic mutagenized. Preferably the sequences are from 50 bp to 100 kb.

The pool of related substrates are converted into overlapping fragments, e.g., from about 5 bp to 5 kb or more, as shown herein. Often, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNaseI or RNase digestion, random shearing or partial restriction enzyme digestion.

Alternatively, the conversion of substrates to fragments can be effected by incomplete PCR amplification of substrates or PCR primed from a single primer. Alternatively, appropriate single-stranded fragments can be generated on a nucleic acid synthesizer. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 100, 500 or 1000.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form. Conversion can be effected by heating to about 80C to 100C, more preferably from 90C to 96 C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with single-stranded DNA binding protein or recA protein. Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 4C to 75C, and preferably from 40C to 65C. Renaturation can be accelerated

by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates as shown herein. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, or proofreading polymerases, such as pfu or pwo, and dNTP's (i.e. dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30T (Stemmer, Proc. Natl. Acad. Sci. USA (1994), supra). The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as stochastic &/or non-stochastic mutagenesis of the nucleic acid in vitro. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb, as shown herein. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from stochastic &/or non-stochastic mutagenesis is used to transform host cells, optionally after cloning into a vector.

#### **8.6.1.2.1.1 FULL LENGTH SEQUENCES**

In a variation of in vitro stochastic &/or non-stochastic mutagenesis, subsequences of recombination substrates can be generated by amplifying the full-length sequences

under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. The amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, wherein at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the incompletely extended products anneal to and prime extension on different sequence-related template species.

#### **8.6.1.2.1.2 OVERLAPPING SINGLE STRANDED DNA FRAGMENTS**

In a further variation, at least one cycle of amplification can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, single-stranded DNA fragments of variable length can be generated from a single primer by Vent DNA polymerase on a first DNA template. The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular single-stranded DNA. This results in multiple substitutions of the first template into the second (see Levichkin et al. Mol. Biology 29:572-577 (1995)).

#### **8.6.1.2.1.3 GENE CLUSTERS**

Gene clusters such as those involved in polyketide synthesis (or indeed any multi-enzyme pathways catalyzing is analogous metabolic reactions) can be recombined by recursive sequence recombination even if they lack DNA sequence homology. Homology can be introduced using synthetic oligonucleotides as PCR primers. In addition to the specific sequences for the gene being amplified, all of the primers used to amplify one type of enzyme (for example the acyl carrier protein in polyketide synthesis) are

synthesized to contain an additional sequence of 20-40 bases 51 to the gene (sequence A) and a different 20-40 base sequence 31 to the gene (sequence B). The adjacent gene (in this case the keto- synthase) is amplified using a 51 primer which contains the complementary strand of sequence B (sequence B'), and a 31 primer containing a different 20-40 base sequence (C). Similarly, primers for the next adjacent gene (keto- reductases) contain sequences C' (complementary to C) and D. If 5 different polyketide gene clusters are being stochastic &/or non-stochastic mutagenized, all five acyl carrier proteins are flanked by sequences A and B following their PCR amplification. In this way, small regions of homology are introduced, making the gene clusters into site specific recombination cassettes. Subsequent to the initial amplification of individual genes, the amplified genes can then be mixed and subjected to primerless PCR. Sequence B at the 3' end of all of the five acyl carrier protein genes can anneal with and prime DNA synthesis from sequence BI at the 5' end of all five keto reductase genes. In this way all possible combinations of genes within the cluster can be obtained. Oligonucleotides allow such recombinants to be obtained in the absence of sufficient sequence homology for recursive sequence recombination described above. Only homology of function is required to produce functional gene clusters.

#### 8.6.1.2.1.4 MULTI SUBUNIT ENZYMES

This method is also useful for exploring permutations of any other multi-subunit enzymes. An example of such enzymes composed of multiple polypeptides that have shown novel functions when the subunits are combined in novel ways are dioxygenases. Directed recombination between the four protein subunits of biphenyl and toluene dioxygenases produced functional dioxygenases with increased activity against trichloroethylene (Furukawa et. al. J. Bacteriol. 176: 2121-2123 (1994)). This combination of subunits from the two dioxygenases could also have been produced by cassette-stochastic &/or non-stochastic mutagenesis of the dioxygenases as described above, followed by selection for degradation of trichloroethylene.

In some polyketide synthases, the separate functions of the acyl carrier protein, keto-synthase, keto- reductase, etc. reside in a single polypeptide. In these cases domains within the single polypeptide may be stochastic &/or non-stochastic mutagenized, even if sufficient homology does not exist naturally, by introducing regions of homology as described above for entire genes. In this case, it may not be possible to introduce additional flanking sequences to the domains, due to the constraint of maintaining a continuous open reading frame.

Instead, groups of oligonucleotides are synthesized that are homologous to the 3' end of the first domain encoded by one of the genes to be stochastic &/or non-stochastic mutagenized, and the 5' ends of the second domains encoded by all of the other genes to be stochastic &/or non-stochastic mutagenized together. This is repeated with all domains, thus providing sequences that allow recombination between protein domains while maintaining their order.

#### **8.6.1.2.1.5 CASSETTE-BASED**

The cassette-based recombination method can be combined with recursive sequence recombination by including gene fragments (generated by DNase, physical shearing, DNA stuttering, etc.) for one or more of the genes. Thus, in addition to different combinations of entire genes within a cluster (e.g., for polyketide synthesis), individual genes can be stochastic &/or non-stochastic mutagenized at the same time (e.g., all acyl carrier protein genes can also be provided as fragmented DNA), allowing a more thorough search of sequence space.

#### **8.6.1.2.1.6 IN VITRO WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

The stochastic &/or non-stochastic mutagenesis of large DNA sequences, such as eukaryotic chromosomes, is difficult by prior art in vitro stochastic &/or non-stochastic mutagenesis methods. A method for overcoming this limitation is described herein.

The cells of related eukaryotic species are gently lysed and the intact chromosomes are liberated. The liberated chromosomes are then sorted by FACS or similar method (such as pulse field electrophoresis) with chromosomes of similar size being sequestered together. Each size fraction of the sorted chromosomes generally will represent a pool of analogous chromosomes, for example the Y chromosome of related mammals. The goal is to isolate intact chromosomes that have not been irreversibly damaged.

The fragmentation and stochastic &/or non-stochastic mutagenesis of such large complex pieces of DNA employing DNA polymerases is difficult and would likely introduce an unacceptably high level of random mutations. An alternative approach that employs restriction enzymes and DNA ligase provides a feasible less destructive solution. A chromosomal fraction is digested with one or more restriction enzymes that recognize long DNA sequences (about 15 - 20bp), such as the intron and intein encoded endonucleases (I-Ppo 1, I-Ceu I, PI-Psp 1, PI-Tli 1, PI-Sce I (VDE).

These enzymes each cut, at most, a few times within each chromosome, resulting in a combinatorial mixture of large fragments, each having overhanging single stranded termini that are complementary to other sites cleaved by the same enzyme.

The digest is further modified by very short incubation with a single stranded exonuclease. The polarity of the nuclease chosen is dependent on the single stranded overhang resulting from the restriction enzyme chosen. 5'-3' exonuclease for 3'-overhangs, and 3'-5' exonuclease for 5'overhangs. This digestion results in significantly long regions of ssDNA overhang on each dsDNA termini. The purpose of this incubation is to generate regions of DNA that define specific regions of DNA where recombination can occur. The fragments are then incubated under condition where the ends of the fragments anneal with other fragments having homologous ssDNA termini. Often, the two fragments annealing will have originated from different chromosomes and in the presence of DNA ligase are covalently linked to form a chimeric chromosome. This generates genetic diversity mimicking the crossing over of homologous chromosomes. The complete ligation reaction

will contain a combinatorial mixture of all possible ligations of fragments having homologous overhanging termini. A subset of this population will be complete chimeric chromosomes.

To screen the stochastic &/or non-stochastic mutagenized library, the chromosomes are delivered to a suitable host in a manner allowing for the uptake and expression of entire chromosomes. For example, YACs (yeast artificial chromosomes) can be delivered to eukaryotic cells by protoplast fusion.

Thus, the reassemble library could be encapsulated in liposomes and fused with protoplasts of the appropriate host cell. The resulting transformants would be propagated and screened for the desired cellular improvements. Once an improved population was identified, the chromosomes would be isolated, stochastic &/or non-stochastic mutagenized, and screened recursively.

#### **8.6.1.2.1.7 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF NATURALLY COMPETENT MICROORGANISMS**

Natural competence is a phenomenon observed for some microbial species whereby individual cells take up DNA from the environment and incorporate it into their genome by homologous recombination. *Bacillus subtilis* and *Acetinetobacter* Spp. are known to be particularly efficient at this process. A method for the whole genome stochastic &/or non-stochastic mutagenesis of these and analogous organisms is described employing this process.

One goal of whole genome stochastic &/or non-stochastic mutagenesis is the rapid accumulation of useful mutations from a population of individual strains into one superior strain. If the organisms to be evolved are naturally competent, then a split pooled strategy for the recursive transformation of naturally competent cells with DNA originating from the pool will effect this process. An example procedure is as follows.

A population of naturally competent organisms that demonstrates a variety of useful traits (such as increased protein secretion) is identified. The strains are pooled, and the pool is split. One half of the pool is used as a source of gDNA, while the other is used to generate a pool of naturally competent cells.

The competent cells are grown in the presence of the pooled gDNA to allow DNA uptake and recombination. Cells of one genotype uptake and incorporate gDNA from cells of a different type generating cells having chimeric genomes. The result is a population of cells representing a combinatorial mixture of the genetic variations originating in the original pool. These cells are pooled again and transformed with the same source of DNA again. This process is carried out recursively to increase the diversity of the genomes of cells resulting from transformation. Once sufficient diversity has been generated, the cell population is screened for new chimeric organisms demonstrating desired improvements.

This process is enhanced by increasing the natural competence of the host organism. COMS is a protein that, when expressed in *B. subtilis*, enhances the efficiency of natural competence mediated transformation more than an order of magnitude.

It was demonstrated that approximately 100% of the cells harboring the plasmid pCOMS uptake and recombine genomic DNA fragments into their genomes. In general, approximately 10% of the genome is recombined into any given transformed cell. This observation was demonstrated by the following.

A strain of *B. subtilis* pCOMS auxotrophic for two nutritional markers was transformed with genomic DNA (gDNA) isolated from a prototrophic strain of the same organism. 10% of the cells exposed to the DNA were prototrophic for one of the two nutrient markers. The average size of the DNA strand taken up by *B. subtilis* is approximately 50kb or about 2% of the genome. Thus 1 of every ten cells had recombined a marker that was represented 1 in every fifty molecules of uptaken gDNA. Thus, most of the cells take up and recombine with approximately five 50kb molecules or 10% of the



genome. This method represents a powerful tool for rapidly and efficiently recombining whole microbial genomes.

In the absence of pCOMS, only 0.3% of the cells prepared for natural competency uptake and integrate a specific marker. This suggested that about 15% of the cells actually underwent recombination with a single genomic fragment. Thus, a recursive transformation strategy as described above produces a whole genome stochastic &/or non-stochastic mutagenized library, even in the absence of pCOMS. In the absence of pCOMS, however, the complex genomes will represent a smaller, but still screenable percentage of the transformed or stochastic &/or non-stochastic mutagenized population.

#### 8.6.1.2.1.8 CONGRESSION

Congression is the integration of two independent unlinked markers into a cell. 0.3% of naturally competent *B. subtilis* cells integrate a single marker (described above). Of these, about 10% have taken up an additional marker. Thus, if one selects or screens for the integration of one specific marker, 10% of the resulting population will have integrated another specific marker. This provides a way of enriching for specific integration events.

For example, if one is looking for the integration of a gene for which there is no easy screen or selection, it will exist as 0.3% of the cell population. If the population is first selected for a specific integration event, then the desired integration will be found in 10% of the population. This represents a significant (about 30-fold) enrichment for the desired event. This enrichment is defined as the "congression effect." The congression effect is not influenced by the presence of pCOMS, thus the "pCOMS effect" is simply to increase the percentage of naturally competent cells that are truly naturally competent from about 15% in its absence to 100% in its presence. All competent cells still uptake about the same amount of DNA or about 10% of the *Bacillus* genome.

The congression effect can be used in the following examples to enhance whole genome stochastic &/or non-stochastic mutagenesis as well, as the targeted integration of stochastic &/or non-stochastic mutagenized genes to the chromosome.

#### **8.6.1.2.1.9 B.SUBTILIS STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

A population of *B. subtilis* cells having desired properties are identified, pooled and stochastic &/or non-stochastic mutagenized as described above with one exception: once the pooled population is split, half of the population is transformed with an antibiotic selection marker that is flanked by sequence that targets its integration and disruption of a specific nutritional gene, for example, one involved in amino biosynthesis. Transformants resistant to the drug are auxotrophic for that nutrient. The resistant population is pooled and grown under conditions rendering them naturally competent (or optionally first transformed with pCOMS).

The competent cells are then transformed with gDNA isolated from the original pool, and prototrophs are selected. The prototrophic population will have undergone recombination with genomic fragments encoding a functional copy of the nutritional marker, and thus will be enriched for cells having undergone recombination at other genetic loci by the congression effect.

#### **8.6.1.2.1.10 TARGETING OF GENES AND GENE LIBRARIES TO THE CHROMOSOME**

It is useful to be able to efficiently deliver genes or gene libraries directly to a specific location in a cells chromosome. As above, target cells are transformed with a positive selection marker flanked by sequences that target its homologous recombination into the chromosome. Selected cells harboring the marker are made naturally competent (with or without pCOMS, but preferably the former) and transformed with a mixture of

two sets of DNA fragments. The first set contains a gene or a stochastic &/or non-stochastic mutagenized library of genes each flanked with sequence to target its integration to a specific chromosomal loci. The second set contains a positive selection marker (different from that first integrated into the cells) flanked by sequence that will target its integration and replacement of the first positive selection marker.

Under optimal conditions, the mixture is such that the gene or gene library is in molar excess over the positive selection marker. Transformants are then selected for cells containing the new positive marker. These cells are enriched for cells having integrated a copy of the desired gene or gene library by the congression effect and can be directly screened for cells harboring the gene or gene variants of interest. This process was carried out using PCR fragments <10kb, and it was found that, employing the congression effect, a population can be enriched such that 50% of the cells are congregants. Thus, one in two cells contained a gene or gene variant.

Alternatively, the expression host can be absent of the first positive selection marker, and the competent cells are transformed with a mixture of the target genes and a limiting amount of the first positive selection marker fragment. Cells selected for the positive marker are screened for the desired properties in the targeted genes. The improved genes are amplified by the PCR, stochastic &/or non-stochastic mutagenized again, and then returned to the original host again with the first positive selection marker. This process is carried out recursively until the desired function of the genes are obtained. This process obviates the need to construct a primary host strain and the need for two positive markers.

#### **8.6.1.2.1.11 CONJUGATION-MEDIATED GENETIC EXCHANGE**

Conjugation can be employed in the evolution of cell genomes in several ways. Conjugative transfer of DNA occurs during contact between cells. See Guiney(1993)in: Bacterial Conjugation (Clewel, ed., Plenum Press, New York), pp. 75-104; Reimann &

Haas in *Bacterial Conjugation* (Clewett, ed., Plenum Press, New York 1993), at pp. 137-188 (incorporated by reference in their entirety for all purposes). Conjugation occurs between many types of gram negative bacteria, and some types of gram positive bacteria. Conjugative transfer is also known between bacteria and plant cells (*Agrobacterium tumefaciens*) or yeast. As discussed in patent 5,837,458, the genes responsible for conjugative transfer can themselves be evolved to expand the range of cell types (e.g., from bacteria to mammals) between which such transfer can occur.

Conjugative transfer is effected by an origin of transfer (*oriT*) and flanking genes (MOB A, B and C, and 15-25 genes, termed *tra*, encoding the structures and enzymes necessary for conjugation to occur. The transfer origin is defined as the site required in *cis* for DNA transfer. *Tra* genes include *tra* A, B, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, U, V, W, X, Y, Z, *virAB* (alleles I-II), C, D, E, G, *IHF*, and *FinOP*. *Tra* genes can be expressed in *cis* or *trans* to *oriT*. Other cellular enzymes, including those of the RecBCD pathway, RecA, SSB protein, DNA gyrase, DNA polII, and DNA ligase, are also involved in conjugative transfer. RecE or recF pathways can substitute for RecBCD.

One structural protein encoded by a *tra* gene is the sex pilus, a filament constructed of an aggregate of a single polypeptide protruding from the cell surface. The sex pilus binds to a polysaccharide on recipient cells and forms a conjugative bridge through which DNA can transfer. This process activates a site-specific nuclease encoded by a MOB gene, which specifically cleaves DNA to be transferred at *oriT*. The cleaved DNA is then threaded through the conjugation bridge by the action of other *tra* enzymes.

Mobilizable vectors can exist in episomal form or integrated into the chromosome. Episomal mobilizable vectors can be used to exchange fragments inserted into the vectors between cells. Integrated mobilizable vectors can be used to mobilize adjacent genes from the chromosome.

#### 8.6.1.2.1.12 USE OF INTEGRATED MOBILIZED VECTORS TO PROMOTE EXCHANGE OF GENOMIC DNA

The F plasmid of *E. coli* integrates into the chromosome at high frequency and mobilizes genes unidirectional from the site of integration (Clewell, 1993, *supra*; Firth et al., in *Escherichia coli and Salmonella Cellular and Molecular Biology* 2, 23 77-2401 (1996); Frost et al., *Microbiol. Rev.* 58, 162-210 (1994)). Other mobilizable vectors do not spontaneously integrate into a host chromosome at high efficiency, but can be induced to do so by growth under particular conditions (e.g., treatment with a mutagenic agent, growth at a nonpermissive temperature for plasmid replication). See Reimann & Haas in *Bacterial Conjugation* (ed. Clewell, Plenum Press, NY 1993), Ch. 6. Of particular interest is the IncP group of conjugal plasmids which are typified by their broad host range (Clewell, 1993, *supra*. Donor "male" bacteria which bear a chromosomal insertion of a conjugal plasmid, such as the *E. coli* F factor can efficiently donate chromosomal DNA to recipient "female" enteric bacteria which lack F (F<sup>-</sup>). Conjugal transfer from donor to recipient is initiated at *oriT*. Transfer of the nicked single strand to the recipient occurs in a 5' to 3' direction by a rolling circle mechanisms which allows mobilization of tandem chromosomal copies. Upon entering the recipient, the donor strand is discontinuously replicated. The linear, single-stranded donor DNA strand is a potent substrate for initiation of *recA*-mediated homologous recombination within the recipient. Recombination between the donor strand and recipient chromosomes can result in the inheritance of donor traits. Accordingly, strains which bear a chromosomal copy of F are designated Hfr (for high frequency of recombination) (Low, 1996 in *Escherichia coli and Salmonella Cellular and Molecular Biology* Vol. 2, pp. 2402- 2405; Sanderson, in *Escherichia coli and Salmonella Cellular and Molecular Biology* 2, 2406-2412 (1996)).

The ability of strains with integrated mobilizable vector to transfer chromosomal DNA provides a rapid and efficient means of exchanging genetic material between a population of bacteria thereby allowing combination of positive mutations and dilution of negative mutations. Such stochastic &/or non-stochastic mutagenesis methods typically start with a population of strains with an integrated mobilizable vector encompassing at least some genetic diversity.

The genetic diversity can be the result of natural variation, exposure to a mutagenic agent or introduction of a fragment library. The population of cells is cultured without selection to allow genetic exchange, recombination and expression of recombinant genes. The cells are then screened or selected for evolution toward a desired property. The population surviving selection/screening can then be subject to a further round of stochastic &/or non-stochastic mutagenesis by IVR-mediated genetic exchange, or otherwise.

The natural efficiency of Hfr and other strains with integrated mob vectors as recipients of conjugal transfer can be improved by several means. The relatively low recipient efficiency of natural BFR strains is attributable to the products of *traS* and *traT* genes of F (Clewell, 1993, *supra*; Firth et al., 1996, *supra*.- Frost et al., 1994, *supra*; Achtman et al., J Mol. Biol. 138, 779-795 (1980). These products are localized to the inner and outer membranes of F+ strains, respectively, where they serve to inhibit redundant matings between two strains which are both capable of donating DNA. The effects of *traS* and *traT*, and cognate genes in other conjugal plasmids, can be eliminated by use of knockout cells incapable of expressing these enzymes or reduced by propagating cells on a carbon- limited source. (Peters et al., J Bacteriol., 178, 3037-3043 (1996)).

In some methods, the starting population of cells has a mobilizable vector integrated at different genomic sites. Directional transfer from *oriT* typically results in more frequent inheritance of traits proximal to *oriT*. This is because mating pairs are fragile and tend to dissociate (particularly when in liquid medium) resulting in the interruption of transfer.

In a population of cells having a mobilizable vector integrated at different sites, chromosomal exchange occurs in a more random fashion. Kits of Hfr strains are available from the E coli. Genetic Stock Center and the Salmonella Genetic Stock Centre (Frost et al., 1994, *supra*).

Alternatively, a library of strains with oriT at random sites and orientations can be produced by insertion mutagenesis using a transposon which bears oriT. The use of a transposon bearing an oriT [e.g., the Tn5-oriT described by Yakobson EA, et al. J. Bacteriol. 1984 Oct; 160(1): 451-453] provides a quick method of generating such a library. Transfer functions for mobilization from the transposon-borne oriT sites are provided by a helper vector in trans. It is possible to generate similar genetic constructs using other sequences known to one of skill as well.

In one aspect, a recursive scheme for genomic stochastic &/or non-stochastic mutagenesis using Tn-oriT elements is provided. A prototrophic bacterial strain or set of related strains bearing a conjugal plasmid, such as the F fertility factor or a member of the IncP group of broad host range plasmids is mutagenized and screened for the desired properties. Individuals with the desired properties are mutagenized with a Tn-oriT element and screened for acquisition of an auxotrophy (e.g., by replica-plating to a minimal and complete media) resulting from insertion of the Tn-oriT element in any one of many biosynthetic gene scattered across the genome. The resulting auxotrophs are pooled and allowed to mate under conditions promoting male-to-male matings, e.g., during growth in close proximity on a filter membrane. Note that transfer functions are provided by the helper conjugal plasmid present in the original strain set. Recombinant transconjugants are selected on minimal medium and screened for further improvement.

Optionally, strains bearing integrated mobilizable vectors are defective in mismatch repair gene(s). Inheritance of donor traits which arise from sequence heterologies increases in strains lacking the methyl-directed mismatch repair system. Optionally, the gene products which decrease recombination efficiency can be inhibited by small molecules.

Intergenic conjugal transfer between species such as *E. coli* and *Salmonella typhimurium*, which are 20% divergent at the DNA level, is also possible if the recipient strain is mutH, mutL or mutS (see Rayssiguier et al., Nature 342, 396-401 (1989)). Such

transfer can be used to obtain recombination at several points as shown by the following example.

One example uses an *S. typhimurium* Hfr donor strain having markers thr557 at map position 0, pyrF2690 at 33 min, serA13 at 62 min and hfrK5 at 43 min. MutS +/-, F- *E. coli*. recipient strains had markers pyrD68 at 21 min aroC355 at 51 min, ilv3164 at 85 min and mutS215 at 59 min. The triauxotrophic *S. typhimurium* Hfr donor and isogenic mutS +/- triauxotrophic *E. coli* recipient were inoculated into 3 ml of Lb broth and shaken at 37C until fully grown. 100 ul of the donor and each recipient were mixed in 10 ml fresh LB broth, and then deposited to a sterile Millipore 0.45 uM HA filter using a Nalgene 250 ml reusable filtration device. The donor and recipients alone were similarly diluted and deposited to check for reversion. The filters with cells were placed cell-side-up on the surface of an LB agar plate which was incubated overnight at 37C. The filters were removed with the aid of a sterile forceps and placed in a sterile 50 ml tube containing 5 ml of minimal salts broth. Vigorous vortexing was used to wash the cells from the filters. 100 ul of mating mixtures, as well as donor and recipient controls were spread to LB for viable cell counts and minimal glucose supplemented with either two of the three recipient requirements for single recombinant counts, one of the three requirements for double recombinant counts, or none of the three requirements for triple recombinant counts. The plates were incubated for 48 hr at 37C after which colonies were counted.

Frequencies are further enhanced by increasing the ratio of donor to recipient cells, or by repeatedly mating the original donor strains with the previously generated recombinant progeny.

#### 8.6.1.2.1.13 INTRODUCTION OF FRAGMENTS BY CONJUGATION

Sobilizable vectors can also be used to transfer fragment libraries into cells to be evolved. This approach is particularly useful in situations in which the cells to be evolved cannot be efficiently transformed directly with the fragment library but can undergo conjugation with primary cells that can be transformed with the fragment library. DNA



fragments to be introduced into host cells encompasses diversity relative to the host cell genome. The diversity can be the result of natural diversity or mutagenesis.

The DNA fragment library is cloned into a mobilizable vector having an origin of transfer. Some such vectors also contain mob genes although alternatively these functions can also be provided in trans. The vector should be capable of efficient conjugal transfer between primary cells and the intended host cells. The vector should also confer a selectable phenotype. This 96 phenotype can be the same as the phenotype being evolved or can be conferred by a marker, such as a drug resistance marker. The vector should preferably allow self-elimination in the intended host cells thereby allowing selection for cells in which a cloned fragment has undergone genetic exchange with a homologous host segment rather than duplication. Such can be achieved by use of vector lacking an origin of replication functional in the intended host type or inclusion of a negative selection marker in the vector.

One suitable vector is the broad host range conjugation plasmid described by Simon et al., *Bio/Technology* 1, 784-791 (1983); Trieu-Cuot et al., *Gene* 102, 99-104 (1991); Bierman et al., *Gene* 116, 43-49 (1992). These plasmids can be transformed into *E. coli* and then force-mated into bacteria that are difficult or impossible to transform by chemical or electrical induction of competence. These plasmids contain the origin of the IncP plasmid, oriT Mobilization functions are supplied in trans by chromosomally-integrated copies of the necessary genes. Conjugal transfer of DNA can in some cases be assisted by treatment of the recipient (if gram-positive) with sub-inhibitory concentrations of penicillins (Trieu-Cuot et al., 1993 *FEMS Microbiol. Lett.* 109, 19-23). To increase diversity in populations, recursive conjugal mating prior to screening is performed.

Cells that have undergone allelic exchange with library fragments can be screened or selected for evolution toward a desired phenotype. Subsequent rounds of recombination can be performed by repeating the conjugal transfer step. The library of fragments can be fresh or can be obtained from some (but not all) of the cells surviving a previous round of selection/screening. Conjugation-mediated stochastic &/or non-stochastic mutagenesis can

be combined with other methods of stochastic &/or non-stochastic mutagenesis.

#### **8.6.1.2.1.14 GENETIC EXCHANGE PROMOTED BY TRANSDUCING PHAGE IN CELLS SUSEPTIBLE TO PHAGE**

Phage transduction can include the transfer, from one cell to another, of nonviral genetic material within a viral coat (Masters, in *Escherichia coli* and *Salmonella Cellular and Molecular Biology* 2, 2421-2442 (1996)). Perhaps the two best examples of generalized transducing phage are bacteriophages P I and P22 of *E. coli* and *S. typhimurium*, respectively. Generalized transducing bacteriophage particles are formed at a low frequency during lytic infection when viral-genome-sized, doubled-stranded fragments of host (which serves as donor) chromosomal DNA are packaged into phage heads. Promiscuous high transducing (HT) mutants of bacteriophage P22 which efficiently package DNA with little sequence specificity have been isolated. Infection of a susceptible host results in a lysate in which up to 50% of the phage are transducing particles. Adsorption of the generalized transducing particle to a susceptible recipient cell results in the injection of the donor chromosomal fragment. RecA-mediated homologous recombination following injection of the donor fragment can result in the inheritance of donor traits. Another type of phage which achieves quasi random insertion of DNA into the host chromosome is Mu. For an overview of Mu biology, see, Groisman (1991) in *Methods in Enzymology* v. 204. Mu can generate a variety of chromosomal rearrangements including deletions, inversions, duplications and transpositions. In addition, elements which combine the features of P22 and Mu are available, including Mud-P22, which contains the ends of the Mu genome in place of the P22 att site and int gene. See, Berg, *supra*.

Generalized transducing phage can be used to exchange genetic material between a population of cells encompassing genetic diversity and susceptible to infection by the phage. Genetic diversity can be the result of natural variation between cells, induced mutation of cells or the introduction of fragment libraries into cells. DNA is then exchanged between cells by generalized transduction. If the phage does not cause lysis of

cells, the entire population of cells can be propagated in the presence of phage. If the phage results in lytic infection, transduction is performed on a split pool basis. That is, the starting population of cells is divided into two. One subpopulation is used to prepare transducing phage. The transducing phage are then infected into the other subpopulation. Preferably, infection is performed at high multiplicity of phage per cell so that few cells remain uninfected. Cells surviving infection are propagated and screened or selected for evolution toward a desired property. The pool of cells surviving screening/selection can then be stochastic &/or non-stochastic mutagenized by a further round of generalized transduction or by other stochastic &/or non-stochastic mutagenesis methods. Recursive split pool transduction is optionally performed prior to selection to increase the diversity of any population to be screened.

The efficiency of the above methods can be increased by reducing infection of cells by infectious (nontransducing phage) and by reducing lysogen formation. The former can be achieved by inclusion of chelators of divalent cations, such as citrate and EDTA in culture media. Tail defective transducing phages can be used to allow only a single round of infection.

Divalent cations are required for phage absorption and the inclusion of chelating agents therefore provides a means of preventing unwanted infection. Integration defective (int) derivatives of generalized transducing phage can be used to prevent lysogen formation. In a further variation, host cells with defects in mismatch repair gene(s) can be used to increase recombination between transduced DNA and genomic DNA.

#### **8.6.1.2.1.15 USE OF LOCKED IN PROPHAGES TO FACILITATE DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

The use of a hybrid, mobile genetic element (locked-in prophages) as a means to facilitate whole genome stochastic &/or non-stochastic mutagenesis of organisms using phage transduction as a means to transfer DNA from donor to recipient is a preferred

embodiment. One such element (Mud-P22) based on the temperate *Salmonella* phage P22 has been described for use in genetic and physical mapping of mutations. See, Youderian et al. (1988) *Genetics* 118:581 - 592, and Benson and Goldman (1992) *J Bacteriol.* 174(5):1673-1681. Individual Mud-P22 insertions package specific regions of the *Salmonella* chromosome into phage P22 particles.

Libraries of random Mud-P22 insertions can be readily isolated and induced to create pools of phage particles packaging random chromosomal DNA fragments. These phage particles can be used to infect new cells and transfer the DNA from the host into the recipient in the process of transduction. Alternatively, the packaged chromosomal DNA can be isolated and manipulated further by techniques such as DNA stochastic &/or non-stochastic mutagenesis or any other mutagenesis technique prior to being reintroduced into cells (especially *recD* cells for linear DNA) by transformation or electroporation, where they integrate into the chromosome. Either the intact transducing phage particles or isolated DNA can be subjected to a variety of mutagens prior to reintroduction into cells to enhance the mutation rate.

Mutator cell lines such as *mutD* can also be used for phage growth. Either method can be used recursively in a process to create genes or strains with desired properties. *E. coli* cells carrying a cosmid clone of *Salmonella* LPS genes are infectable by P22 phage. It is possible to develop similar genetic elements using other combinations of transposable elements and bacteriophages or viruses as well. P22 is a lambdoid phage that packages its DNA into stochastic &/or non-stochastic mutagenized phage particles (heads) by a "headful" mechanism. Packaging of phage DNA is initiated at a specific site (*pac*) and proceeds unidirectionally along a linear, double stranded normally concatameric molecule. When the phage head is full (about 43 kb), the DNA strand is cleaved, and packaging of the next phage head is initiated. Locked-in or excision-defective P22 prophages, however, initiate packaging at their *pac* site, and then proceed unidirectionally along the chromosome, packaging successive headfuls of chromosomal DNA (rather than phage DNA). When these transducing phages infect new *Salmonella* cells they inject the chromosomal DNA from the original host into the recipient cell, where it can recombine

into the chromosome by homologous recombination creating a chimeric chromosome. Upon infection of recipient cells at a high multiplicity of infection, recombination can also occur between incoming transducing fragments prior to recombination into the chromosome.

Integration of such locked-in P22 prophages at various sites in the chromosome allows flanking regions to be amplified and packaged into phage particles. The Mud-P22 mobile genetic element contains an excision-defective P22 prophage flanked by the ends of phage/transposon Mu. The entire Mud-P22 element can transpose to virtually any location in the chromosome or other episome (eg. F', BAC clone) when the Mu A and B proteins are provided in trans.

A number of embodiments for this type of genetic element are available. In one example, the locked in prophage are used as generalized transducing phage to transfer random fragments of a donor chromosome into a recipient. The Mud-P22 element acts as a transposon when Mu A and B transposase proteins are provided in trans and integrate copies of itself at random locations in the chromosome. In this way, a library of random chromosomal Mud-P22 insertions can be generated in a suitable host. When the Mud-P22 prophages in this library are induced, random fragments of chromosomal DNA will be packaged into phage particles. When these phages infect recipient cells, the chromosomal DNA is injected and can recombine into the chromosome of the recipient. These recipient cells are screened for a desired property and cells showing improvement are then propagated.

The process can be repeated, since the Mud-P22 genetic element is not transferred to the recipient in this process. Infection at a high multiplicity allows for multiple chromosomal fragments to be injected and recombined into the recipient chromosome. Locked in prophages can also be used as specialized transducing phage.

Individual insertions near a gene of interest can be isolated from a random insertion library by a variety of methods. Induction of these specific prophages results in

packaging of flanking chromosomal DNA including the gene(s) of interest into phage particles. Infection of recipient cells with these phages and recombination of the packaged DNA into the chromosome creates chimeric genes that can be screened for desired properties. Infection at a high multiplicity of infection can allow recombination between incoming transducing fragments prior to recombination into the chromosome.

These specialized transducing phage can also be used to isolate large quantities of high quality DNA containing specific genes of interest without any prior knowledge of the DNA sequence. Cloning of specific genes is not required. Insertion of such an element nearby a biosynthetic operon for example allows for large amounts of DNA from that operon to be isolated for use in DNA stochastic &/or non-stochastic mutagenesis (in vitro and/or in vivo), cloning, sequencing, or other uses as set forth herein. DNA isolated from similar insertions in other organisms containing homologous operons are optionally mixed for use in family stochastic &/or non-stochastic mutagenesis formats as described, herein, in which homologous genes from different organisms (or different chromosomal locations within a single species, or both). Alternatively, the transduced population is recursively transduced with pooled transducing phage or new transducing phage generated from the previously transduced cells. This can be carried out recursively to optimize the diversity of the genes prior to stochastic &/or non-stochastic mutagenesis.

Phage isolated from insertions in a variety of strains or organisms containing homologous operons are optionally mixed and used to coinfect cells at a high MOI allowing for recombination between incoming transducing fragments prior to recombination into the chromosome.

Locked in prophage are useful for mapping of genes, operons, and/or specific mutations with either desirable or undesirable phenotypes. Locked-in prophages can also provide a means to separate and map multiple mutations in a given host. If one is looking for beneficial mutations outside a gene or operon of interest, then an unmodified gene or operon can be transduced into a mutagenized or stochastic &/or non-stochastic mutagenized host then screened for the presence of desired secondary mutations.

Alternatively, the gene/operon of interest can be readily moved from a mutagenized/stochastic &/or non-stochastic mutagenized host into a different background to screen/select for modifications in the gene/operon itself. It is also possible to develop similar genetic elements using other combinations of transposable elements and bacteriophages or viruses as well. Similar systems are set up in other organisms, e.g., that do not allow replication of P22 or P1. Broad host range phages and transposable elements are especially useful. Similar genetic elements are derived from other temperate phages that also package by a headful mechanism. In general, these are the phages that are capable of generalized transduction. Viruses infecting eukaryotic cells may be adapted for similar purposes. Examples of generalized transducing phages that are useful are described in: Green et al., "Isolation and preliminary characterization of lytic and lysogenic phages with wide host range within the streptomycetes", *J Gen Microbiol* 131(9):2459-2465 (1985); Studdard et al., "Genome structure in *Streptomyces* spp.: adjacent genes on the *S. coelicolor* A3(2) linkage map have cotransducible analogs in *S. venezuelae*", *J Bacteriol* 169(8):3 814-3 816 (1987); Wang et al., "High frequency generalized transduction by miniMu plasmid phage", *Genetics* 116(2):201-206, (1987); Welker, N. E., "Transduction in *Bacillus stearothermophilus*", *J Bacteriol*, 176(11):3354-3359, (1988); Darzins et al., "Mini-D3112 bacteriophage transposable elements for genetic analysis of *Pseudomonas aeruginosa*", *J Bacteriol* 171(7):3909-3916 (1989); Hugouvieux-Cotte-Pattat et al., "Expanded linkage map of *Erwinia chrysanthemi* strain 3937", *Mol Microbiol* 3(5):573-581, (1989); Ichige et al. "Establishment of gene transfer systems for and construction of the genetic map of a marine *Vibrio* strain", *J Bacteriol* 171(4):1825-1834 (1989); Murainatsu et al., "Two generalized transducing phages in *Vibrio parahaemolyticus* and *Vibrio alginolyticus*", *Microbiol Immunol* (12):1073-1084 (1991); Regue et al., "A generalized transducing bacteriophage for *Serratia marcescens*", *Res Microbiol* 42(1):23 - 27, (199 1) - Kiesel et al , "Phage Acn I -mediated transduction in the facultatively methanol-utilizing *A. cetobacter methanolicus* MB 58/4", *J Gen Virol* 74(9):1741-1745 (1993); Blahova et al., "Transduction of imipenem resistance by the phage F- 116 from a nosocomial strain of *Pseudomonas aeruginosa* isolated in Slovakia", *Acta Virol* 38(5):247-250 (1994); Kidambi et al., "Evidence for phage- mediated gene transfer among *Pseudomonas aeruginosa* strains on the phylloplane", *Appl Environ Microbiol* 60:(2)496-

500 (1994); Weiss et al., "Isolation and characterization of a generalized transducing phage for *Xanthomonas campestris* pv. *campestris*", *J Bacteriol* 176(11):3354-3359 (1994); Matsumoto et al., "Clustering of the *trp* genes in *Burkholderia* (formerly *Pseudomonas*) *cepacia*", *FEMS Microbiol Lett* 134(2-3):265-271 (1995); Schicklmaier et al., "Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex", *Appl Environ Microbiol* 61(4):1637-1640 (1995); Humphrey et al., "Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*", *J Bacteriol* 179(2):323-329 (1997); Willi et al., "Transduction of antibiotic resistance markers among *Actinobacillus actinomycetemcomitans* strains by temperate bacteriophages  $\lambda$  phi 23", *Cell Mol Life Sci* 53 (11-12):904-910 (1997); Jensen et al., "Prevalence of broad-host-range lytic bacteriophages of *Sphaerofilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*", *Appl Environ Microbiol* 64(2):575-580 (1998), and Nedelmann et al., "Generalized transduction for genetic linkage analysis and transfer of transposon insertions in different *Staphylococcus epidermidis* strains", *Zentralblatt Bakteriologie* 287(1-2):85-92 (1998).

A Mud-PI/Tn-P1 system comparable to Mud-P22 is developed using phage P1. Phage P1 has an advantage of packaging much larger (about 110 kb) fragments per headful. Phage P1 is currently used to create bacterial artificial chromosomes or BAC's. P1-based BAC vectors are designed along these principles so that cloned DNA is packaged into phage particles, rather than the current system, which requires DNA preparation from single-copy episomes. This combines the advantages of both systems in having the genes cloned in a stable single-copy format, while allowing for amplification and specific packaging of cloned DNA upon induction of the prophage.

#### **8.6.1.2.1.16 RANDOM PLACEMENT OF GENES OR IMPROVED GENES THROUGHOUT THE GENOME FOR OPTIMIZATION OF GENE CONTEXT**

The placement and orientation of genes in a host chromosome (the "context" of the gene in a chromosome) or episome has large effects on gene expression and activity. Random integration of plasmid or other episomal sequences into a host chromosome by



non-homologous recombination, followed by selection or screening for the desired phenotype, is a preferred way of identifying optimal chromosomal positions for expression of a target. This strategy is illustrated herein.

A variety of transposon mediated delivery systems can be employed to deliver genes of interest, either individual genes, genomic libraries, or a library of stochastic &/or non-stochastic mutagenized gene(s) randomly throughout the genome of a host. Thus, in one preferred embodiment, the improvement of a cellular function is achieved by cloning a gene of interest, for example a gene encoding a desired metabolic pathway, within a transposon delivery vehicle.

Such transposon vehicles are available for both Gram-negative and Gram-positive bacteria. De Lorenzo and Timis (1994) *Methods in Enzymology* 235:385-404 describe the analysis and construction of stable phenotypes in gram-negative Bacteria with Tn5- and Tn 10-derived minitransposons. Kleckner et al. (1991) *Methods in Enzymology* 204, chapter 7 describe uses of transposons such as Tn 10, including for use in gram positive bacteria. Petit et al. (1990) *Journal of Bacteriology*, 172(12):6736-6740 describe Tn10 derived transposons active in *Bacillus Subtilis*. The transposon delivery vehicle is introduced into a cell population, which is then selected for recombinant cells that have incorporated the transposon into the genome.

The selection is typically by any of a variety of drug resistant markers also carried within the transposon. The selected subpopulation is screened for cells having improved expression of the gene(s) of interest. Once cells harboring the genes of interest in the optimal location are isolated, the genes are amplified from within the genome using PCR, stochastic &/or non-stochastic mutagenized, and cloned back into a similar transposon delivery vehicle which contains a different selection marker within the transposon and lacks the transposon integrase gene.

This stochastic &/or non-stochastic mutagenized library is then transformed back into the strain harboring the original transposon, and the cells are selected for the presence

of the new resistance marker and the loss of the previous selection marker. Selected cells are enriched for those that have exchanged by homologous recombination the original transposon for the new transposon carrying members of the stochastic &/or non-stochastic mutagenized library. The surviving cells are then screened for further improvements in the expression of the desired phenotype. The genes from the improved cells are then amplified by the PCR and stochastic &/or non-stochastic mutagenized again. This process is carried out recursively, oscillating each cycle between the different selection markers. Once the gene(s) of interest are optimized to a desired level, the fragment can be amplified and again randomly distributed throughout the genome as described above to identify the optimal location of the improved genes.

Alternatively, the gene(s) conferring a desired property may not be known. In this case the DNA fragments cloned within the transposon delivery vehicle could be a library of genomic fragments originating from a population of cells derived from one or more strains having the desired property(ies). The library is delivered to a population of cells derived from one or more strains having or lacking the desired property(ies) and cells incorporating the transposon are selected. The surviving cells are then screened for acquisition or improvement of the desired property. The fragments contained within the surviving cells are amplified by PCR and then cloned as a pool into a similar transposon delivery vector harboring a different selection marker from the first delivery vector. This library is then delivered to the pool of surviving cells, and the population having acquired the new selective marker is selected. The selected cells are then screened for further acquisition or improvement of the desired property.

In this way the different possible combinations of genes conferring or improving a desired phenotype are explored in a combinatorial fashion. This process is carried out repetitively with each new cycle employing an additional selection marker. Alternatively, PCR fragments are cloned into a pool of transposon vectors having different selective markers. These are delivered to cells and selected for 1, 2, 3, or more markers.

Alternatively, the amplified fragments from each improved cell are stochastic &/or

non-stochastic mutagenized independently. The stochastic &/or non-stochastic mutagenized libraries are then cloned back into a transposon delivery vehicle similar to the original vector but containing a different selection marker and lacking the transposase gene. Selection is then for acquisition of the new marker and loss of the previous marker. Selected cells are enriched for those incorporating the stochastic &/or non-stochastic mutagenized variants of the amplified genes by homologous recombination. This process is carried out recursively, oscillating each cycle between the two selective markers.

#### **8.6.1.2.1.17 IMPROVEMENT OF OVEREXPRESSED GENES FOR A DESIRED PHENOTYPE**

The improvement of a cellular property or phenotype is often enhanced by increasing the copy number or expression of gene(s) participating in the expression of that property. Genes that have such an effect on a desired property can also be improved by DNA stochastic &/or non-stochastic mutagenesis to have a similar effect. A genomic DNA library is cloned into an overexpression vector and transformed into a target cell population such that the genomic fragments are highly expressed in cells selected for the presence of the overexpression vector. The selected cells are then screened for improvement of a desired property. The overexpression vector from the improved cells are isolated and the cloned genomic fragments stochastic &/or non-stochastic mutagenized. The genomic fragment carried in the vector from each improved isolate is stochastic &/or non-stochastic mutagenized independently or with identified homologous genes (family stochastic &/or non-stochastic mutagenesis). The stochastic &/or non-stochastic mutagenized libraries are then delivered back to a population of cells and the selected transformants rescreened for further improvements in the desired property. This stochastic &/or non-stochastic mutagenesis/screening process is cycled recursively until the desired property has been optimized to the desired level. As stated above, gene dosage can greatly enhance a desired cellular property.

One method of increasing gene copy number of unknown genes is using a method of random amplification (see also, Mavingui et. al. (1997) Nature Biotech, 15, 5 64). In

this method, a genomic library is cloned into a suicide vector containing a selective marker that also at higher dosage provides an enhanced phenotype. An example of such a marker is the kanamycin resistance gene. At successively higher copy number, resistance to successively higher levels of kanamycin is achieved. The genomic library is delivered to a target cell by any of a variety of methods including transformation, transduction, conjugation, etc. Cells that have incorporated the vector into the chromosome by homologous recombination between the vector and chromosomal copies of the cloned genes can be selected by requiring expression of the selection marker under conditions where the vector does not replicate. This recombination event results in the duplication of the cloned DNA fragment in the host chromosome with a copy of the vector and selection marker separating the two copies. The population of surviving cells are screened for improvement of a desired cellular property resulting from the gene duplication event. Further gene duplication events resulting in additional copies of the original cloned DNA fragments can be generated by further propagating the cells under successively more stringent selective conditions i.e. increased concentrations of kanamycin. In this case selection requires increased copies of the selective marker, but increased copies of the desired gene fragment is also concomitant. Surviving cells are further screened for an improvement in the desired phenotype. The resulting population of cells likely resulted in the amplification of different genes since often many genes effect a given phenotype. To generate a library of the possible combinations of these genes, the original selected library showing phenotypic improvements are recombined, using the methods described herein, e.g., protoplast fusion, split pool transduction, transformation, conjugation, etc.

The recombined cells are selected for increased expression of the selective marker. Survivors are enriched for cells having incorporated additional copies of the vector sequence by homologous recombination, and these cells will be enriched for those having combined duplications of different genes. In other words, the duplication from one cell of enhanced phenotype becomes combined with the duplication of another cell of enhanced phenotype. These survivors are screened for further improvements in the desired phenotype. This procedure is repeated recursively until the desired level of phenotypic expression is achieved.

Alternatively, genes that have been identified or are suspected as being beneficial in increased copy number are cloned in tandem into appropriate plasmid vectors. These vectors are then transformed and propagated in an appropriate host organism. Plasmid-plasmid recombination between the cloned gene fragments result in further duplication of the genes. Resolution of the plasmid doublet can result in the uneven distribution of the gene copies, with some plasmids having additional gene copies and others having fewer gene copies. Cells carrying this distribution of plasmids are then screened for an improvement in the phenotype effected by the gene duplications.

In summary, a method of selecting for increased copy number of a nucleic acid sequence by the above procedure is provided. In the method, a genomic library in a suicide vector comprising a dose-sensitive selectable marker is provided, as noted above. The genomic library is transduced into a population of target cells. The target cells are selected in a population of target cells for increasing doses of the selectable marker under conditions in which the suicide vector does not replicate episomally. A plurality of target cells are selected for the desired phenotype, recombined and reselected. The process is recursively repeated, if desired, until the desired phenotype is obtained.

#### **8.6.1.2.1.18 STRATEGIES FOR IMPROVING GENOMIC STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS VIA TRANSFORMATION OF LINEAR DNA FRAGMENTS**

Wild-type members of the Enterobacteriaceae (e.g., *Escherichia coli*) are typically resistant to genetic exchange following transformation of linear DNA molecules.

This is due, at least in part, to the Exonuclease V (Exo V) activity of the RecBCD holoenzyme which rapidly degrades linear DNA molecules following transformation. Production of ExoV has been traced to the *recD* gene, which encodes the D subunit of the holoenzyme. As demonstrated by Russel et al. (1989) *Journal of Bacteriology* 2609-2613,

homologous recombination between a transformed linear donor DNA molecule and the chromosome of recipient is readily detected in a strains bearing a loss of function mutation in a *recD* mutant.

The use of *recD* strains provides a simple means for genomic stochastic &/or non-stochastic mutagenesis of the Enterobacteriaceae. For example, a bacterial strain or set of related strains bearing a *recD* null mutation (e.g., the *E. coli recD1903::mini-Tet* allele) is mutagenized and screened for the desired properties. In a split-pool fashion, chromosomal DNA prepared on one aliquot could be used to transform (e.g., via electroporation or chemically induced competence) the second aliquot. The resulting transformants are then screened for improvement, or recursively transformed prior to screening.

The use of *RecE/ recT* as described supra, can improve homologous recombination of linear DNA fragments. The *RecBCD* holoenzyme plays an important role in initiation of *RecA*-dependent homologous recombination. Upon recognizing a dsDNA end, the *RecBCD* enzyme unwinds and degrades the DNA asymmetrically in a 5' to 3' direction until it encounters a chi (or 'X')-site (consensus 5'-GCTGGTGG-3') which attenuates the nuclease activity. This results in the generation of a ssDNA terminating near the c site with a 3'-ssDNA tail that is preferred for *RecA* loading and subsequent invasion of dsDNA for homologous recombination. Accordingly, preprocessing of transforming fragments with a 5' to 3' specific ssDNA Exonuclease, such as *Lamda* ( ) exonuclease (available, e.g., from Boeringer Mannheim) prior to transformation may serve to stimulate homologous recombination in *recD*- strain by providing ssDNA invasive end for *RecA* loading and subsequent strand invasion.

The addition of DNA sequence encoding chi-sites (consensus 5'-GCTGGTGG-3') to DNA fragments can serve to both attenuate Exonuclease V activity and stimulate homologous recombination, thereby obviating the need for a *recD* mutation (see also, Kowalczykowski, et al. (1994) "Biochemistry of homologous recombination in *Escherichia coli*," *Microbiol. Rev.* 58:401-465 and Jessen, et al. (1998) "Modification of bacterial artificial chromosomes through Chi-stimulated homologous recombination and

its application in zebrafish transgenesis." *Proc. Natl. Acad. Sci.* 95:5121- 5126). Chi sites are optionally included in linkers ligated to the ends of transforming fragments or incorporated into the external primers used to generate DNA fragments to be transformed. The use of recombination-stimulatory sequences such as chi is a generally useful approach for evolution of a broad range of cell types by fragment transformation. Methods to inhibit or mutate analogs of Exo V or other nucleases (such as, Exonucleases I (endA 1), 111 (nth), IV (nfo), VII, and VIII of *E. coli*) is similarly useful.

Inhibition or elimination of nucleases, or modification of ends of transforming DNA fragments to render them resistant to exonuclease activity has applications in evolution of a broad range of cell types.

#### **8.6.1.2.1.19 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS TO OPTIMIZE UNKNOWN INTERACTIONS**

Many observed traits are the result of complex interactions of multiple genes or gene products. Most such interactions are still uncharacterized. Accordingly, it is often unclear which genes need to be optimized to achieve a desired trait, even if some of the genes contributing to the trait are known.

This lack of characterization is not an issue during DNA stochastic &/or non-stochastic mutagenesis, which produces solutions that optimize whatever is selected for. An alternative approach, which has the potential to solve not only this problem, but also anticipated future rate limiting factors, is complementation by overexpression of unknown genomic sequences.

A library of genomic DNA is first made as described, supra. This is transformed into the cell to be optimized and transformants are screened for increases in a desired property. Genomic fragments which result in an improved property are evolved by DNA stochastic &/or non-stochastic mutagenesis to further increase their beneficial effect. This

approach requires no sequence information, nor any knowledge or assumptions about the nature of protein or pathway interactions, or even of what steps are rate-limiting; it relies only on detection of the desired phenotype. This sort of random cloning and subsequent evolution by DNA stochastic &/or non-stochastic mutagenesis of positively interacting genomic sequences is extremely powerful and generic. A variety of sources of genomic DNA are used, from isogenic strains to more distantly related species with potentially desirable properties. In addition, the technique is applicable to any cell for which the molecular biology basics of transformation and cloning vectors are available, and for any property which can be assayed (preferably in a high-throughput format). Alternatively, once optimized, the evolved DNA can be returned to the chromosome by homologous recombination or randomly by phage mediated site-specific recombination.

#### **8.6.1.2.1.20 HOMOLOGOUS RECOMBINATION WITHIN THE CHROMOSOME**

Homologous recombination within the chromosome is used to circumvent the limitations of plasmid based evolution and size restrictions. The strategy is similar to that described above for stochastic &/or non-stochastic mutagenesis genes within their chromosomal context, except that no in vitro stochastic &/or non-stochastic mutagenesis occurs. Instead, the parent strain is treated with mutagens such as ultraviolet light or nitrosoguanidine, and improved mutants are selected. The improved mutants are pooled and split. Half of the pool is used to generate random genomic fragments for cloning into a homologous recombination vector. Additional genomic fragments are optionally derived from related species with desirable properties. The cloned genomic fragments are homologously recombined into the genomes of the remaining half of the mutant pool, and variants with improved properties are selected. These are subjected to a further round of mutagenesis, selection and recombination. Again this process is entirely generic for the improvement of any whole cell biocatalyst for which a recombination vector and an assay can be developed. Here again, it should be noted that recombination can be performed recursively prior to screening.



#### **8.6.1.2.1.21 METHODS FOR RECURSIVE SEQUENCE RECOMBINATION**

As shown herein, DNA Stochastic &/or non-stochastic mutagenesis provides most rapid technology for evolution of complex new functions. As shown herein, recombination in DNA stochastic &/or non-stochastic mutagenesis achieves accumulation of multiple beneficial mutations in a few cycles. In contrast, because of the high frequency of deleterious mutations relative to beneficial ones, iterative point mutation must build beneficial mutations one at a time, and consequently requires many cycles to reach the same point. As shown herein, rather than a simple linear sequence of mutation accumulation, DNA stochastic &/or non-stochastic mutagenesis is a parallel process where multiple problems may be solved independently, and then combined.

#### **8.6.1.2.2 IN VIVO FORMATS**

##### **8.6.1.2.2.1 PLASMID-PLASMID RECOMBINATION**

The initial substrates for recombination are a collection of polynucleotides comprising variant forms of a gene. The variant forms usually show substantial sequence identity to each other sufficient to allow homologous recombination between substrates. The diversity between the polynucleotides can be natural (e.g., allelic or species variants), induced (e.g., error-prone PCR or error-prone recursive sequence recombination), or the result of in vitro recombination. Diversity can also result from resynthesizing genes encoding natural proteins with alternative codon usage. There should be at least sufficient diversity between substrates that recombination can generate more diverse products than there are starting materials. There must be at least two substrates differing in at least two positions.

However, commonly a library of substrates of  $10^3$ - $10^8$  members is employed. The degree of diversity depends on the length of the substrate being recombined and the extent of the functional change to be evolved. Diversity at between 0.1-25% of positions is typical. The diverse substrates are incorporated into plasmids. The plasmids are often standard cloning vectors, e.g., bacterial multicopy plasmids. However, in some methods to

be described below, the plasmids include mobilization (MOB) functions. The substrates can be incorporated into the same or different plasmids. Often at least two different types of plasmid having different types of selectable markers are used to allow selection for cells containing at least two types of vector. Also, where different types of plasmid are employed, the different plasmids can come from two distinct incompatibility groups to S allow stable co-existence of two different plasmids within the cell. Nevertheless, plasmids from the same incompatibility group can still co-exist within the same cell for sufficient time to allow homologous recombination to occur.

Plasmids containing diverse substrates are initially introduced into cells by any method (e.g., chemical transformation, natural competence, electroporation, biolistics, packaging into phage or viral systems). Often, the plasmids are present at or near saturating concentration (with respect to maximum transfection capacity) to increase the probability of more than one plasmid entering the same cell. The plasmids containing the various substrates can be transfected simultaneously or in multiple rounds. For example, in the latter approach cells can be transfected with a first aliquot of plasmid, transfectants selected and propagated, and then infected with a second aliquot of plasmid.

Having introduced the plasmids into cells, recombination between substrates to generate recombinant genes occurs within cells containing multiple different plasmids merely by propagating the cells. However, cells that receive only one plasmid are unable to participate in recombination and the potential contribution of substrates on such plasmids to evolution is not fully exploited (although these plasmids may contribute to some extent if they are propagated in mutator cells). The rate of evolution can be increased by allowing all substrates to participate in recombination. Such can be achieved by subjecting transfected cells to electroporation. The conditions for electroporation are the same as those conventionally used for introducing exogenous DNA into cells (e.g., 1,000-2,500 volts, 400 uF and a 1-2 mM gap). Under these conditions, plasmids are exchanged between cells allowing all substrates to participate in recombination.

In addition the products of recombination can undergo further rounds of recombination with each other or with the original substrate. The rate of evolution can also be increased by use of conjugative transfer. To exploit conjugative transfer, substrates can be cloned into plasmids having MOB genes, and tra genes are also provided in cis or in trans to the MOB genes. The effect of conjugative transfer is very similar to electroporation in that it allows plasmids to move between cells and allows recombination between any substrate and the products of previous recombination to occur, merely by propagating the culture. The rate of evolution can also be increased by fusing cells to induce exchange of plasmids or chromosomes. Fusion can be induced by chemical agents, such as PEG, or viral proteins, such as influenza virus hemagglutinin, HSV-1 gB and gD. The rate of evolution can also be increased by use of mutator host cells (e.g., Mut L, S, D, T, H in bacteria and Ataxia telangiectasia human cell lines) .

The time for which cells are propagated and recombination is allowed to occur, of course, varies with the cell type but is generally not critical, because even a small degree of recombination can substantially increase diversity relative to the starting materials. Cells bearing plasmids containing recombined genes are subject to screening or selection for a desired function. For example, if the substrate being evolved contains a drug resistance gene, one would select for drug resistance. Cells surviving screening or selection can be subjected to one or more rounds of screening/selection followed by recombination or can be subjected directly to an additional round of recombination. The next round of recombination can be achieved by several different formats independently of the previous round. For example, a further round of recombination can be effected simply by resuming the electroporation or conjugation-mediated intercellular transfer of plasmids described above.

Alternatively, a fresh substrate or substrates, the same or different from previous substrates, can be transfected into cells surviving selection/screening. Optionally, the new substrates are included in plasmid vectors bearing a different selective marker and/or from

a different incompatibility group than the original plasmids. As a further alternative, cells surviving selection/screening can be subdivided into two subpopulations, and plasmid DNA from one subpopulation transfected into the other, where the substrates from the plasmids from the two subpopulations undergo a further round of recombination. In either of the latter two options, the rate of evolution can be increased by employing DNA extraction, electroporation, conjugation or mutator cells, as described above. In a still further variation, DNA from cells surviving screening/selection can be extracted and subjected to *in vitro* recursive sequence recombination. After the second round of recombination, a second round of screening/selection is performed, preferably under conditions of increased stringency. If desired, further rounds of recombination and selection/screening can be performed using the same strategy as for the second round.

With successive rounds of recombination and selection/screening, the surviving recombined substrates evolve toward acquisition of a desired phenotype. Typically, in this and other methods of recursive recombination, the final product of recombination that has acquired the desired phenotype differs from starting substrates at 0.1%-50% of positions and has evolved at a rate orders of magnitude in excess (e.g., by at least 10-fold, 100-fold, 1000-fold, or 10,000 fold) of the rate of naturally acquired mutation of about 1 mutation per  $10^9$  positions per generation (see Anderson et al. Proc. Natl. Acad. Sci. U.S.A. 93:906-907 (1996)). The "final product" may be transferred to another host more desirable for utilization of the "stochastic &/or non-stochastic mutagenized" DNA.

This is particularly advantageous in situations where the more desirable host is less efficient as a host for the many cycles of mutation/recombination due to the lack of molecular biology or genetic tools available for other organisms such as *E. coli*.

#### **8.6.1.2.2.2 VIRUS-PLASMID RECOMBINATION**

The strategy used for plasmid-plasmid recombination can also be used for virus-plasmid recombination; usually, phage-plasmid recombination. However, some additional comments particular to the use of viruses are appropriate.

The initial substrates for recombination are cloned into both plasmid and viral vectors. It is usually not critical which substrate(s) are inserted into the viral vector and which into the plasmid, although usually the viral vector should contain different substrate(s) from the plasmid. As before, the plasmid (and the virus) typically contains a selective marker.

The plasmid and viral vectors can both be introduced into cells by transfection as described above. However, a more efficient procedure is to transfect the cells with plasmid, select transfectants and infect the transfectants with virus. Because the efficiency of infection of many viruses approaches 100% of cells, most cells transfected and infected by this route contain both a plasmid and virus bearing different substrates.

Homologous recombination occurs between plasmid and virus generating both recombined plasmids and recombined virus. For some viruses, such as filamentous phage, in which intracellular DNA exists in both double-stranded and single-stranded forms, both can participate in recombination.

Provided that the virus is not one that rapidly kills cells, recombination can be augmented by use of electroporation or conjugation to transfer plasmids between cells. Recombination can also be augmented for some types of virus by allowing the progeny virus from one cell to reinfect other cells. For some types of virus, virus infected-cells show resistance to superinfection. However, such resistance can be overcome by infecting at high multiplicity and/or using mutant strains of the virus in which resistance to superinfection is reduced.

The result of infecting plasmid-containing cells with virus depends on the nature of the virus. Some viruses, such as filamentous phage, stably exist with a plasmid in the cell and also extrude progeny phage from the cell. Other viruses, such as lambda having a cosmid genome, stably exist in a cell like plasmids without producing progeny virions.

Other viruses, such as the T-phage and lytic lambda, undergo recombination with the plasmid but ultimately kill the host S cell and destroy plasmid DNA. For viruses that infect cells without killing the host, cells containing recombinant plasmids and virus can be screened/selected using the same approach as for plasmid-plasmid recombination. Progeny virus extruded by cells surviving selection/screening can also be collected and used as substrates in subsequent rounds of recombination. For viruses that kill their host cells, recombinant genes resulting from recombination reside only in the progeny virus. If the screening or selective assay requires expression of recombinant genes in a cell, the IS recombinant genes should be transferred from the progeny virus to another vector, e.g., a plasmid vector, and retransfected into cells before selection/screening is performed.

For filamentous phage, the products of recombination are present in both cells surviving recombination and in phage extruded from these cells. The dual source of recombinant products provides some additional options relative to the plasmid-plasmid recombination. For example, DNA can be isolated from phage particles for use in a round of in vitro recombination. Alternatively, the progeny 2S phage can be used to transfect or infect cells surviving a previous round of screening/selection, or fresh cells transfected with fresh substrates for recombination.

#### 8.6.1.2.2.3 VIRUS-VIRUS RECOMBINATION

The principles described for plasmid-plasmid and plasmid-viral recombination can be applied to virus-virus recombination with a few modifications. The initial substrates for recombination are cloned into a viral vector. Usually, the same vector is used for all substrates.

Preferably, the virus is one that, naturally or as a result of mutation, does not kill cells. After insertion, some viral genomes can be packaged in vitro or using a packaging cell line. The packaged viruses are used to infect cells at high multiplicity such that there is a high probability that a cell will receive multiple viruses bearing different substrates.

After the initial round of infection, subsequent steps depend on the nature of infection as discussed in the previous section. For example, if the viruses have phagemid genomes such as lambda cosmids or M13, F1 or Fd phagemids, the phagemids behave as plasmids within the cell and undergo recombination simply by propagating the cells. Recombination is particularly efficient between single-stranded forms of intracellular DNA. Recombination can be augmented by electroporation of cells.

Following selection/screening, cosmids containing recombinant genes can be recovered from surviving cells, e.g., by heat induction of a cos- lysogenic host cell, or extraction of DNA by standard procedures, followed by repackaging cosmid DNA in vitro.

If the viruses are filamentous phage, recombination of replicating form DNA occurs by propagating the culture of infected cells. Selection/screening identifies colonies of cells containing viral vectors having recombinant genes with improved properties, together with phage extruded from such cells. Subsequent options are essentially the same as for plasmid-viral recombination.

#### **8.6.1.2.2.4 CHROMOSOME RECOMBINATION**

This format can be used to especially evolve chromosomal substrates. The format is particularly useful in situations in which many chromosomal genes contribute to a phenotype or one does not know the exact location of the chromosomal gene(s) to be evolved. The initial substrates for recombination are cloned into a plasmid vector. If the chromosomal gene(s) to be evolved are known, the substrates constitute a family of sequences showing a high degree of sequence identity but some divergence from the chromosomal gene. If the chromosomal genes to be evolved have not been located, the initial substrates usually constitute a library of DNA segments of which only a small number show sequence identity to the gene or gene(s) to be evolved. Divergence between plasmid-borne substrate and the chromosomal gene(s) can be induced by mutagenesis or

by obtaining the plasmid- borne substrates from a different species than that of the cells bearing the chromosome.

The plasmids bearing substrates for recombination are transfected into cells having chromosomal gene(s) to be evolved. Evolution can occur simply by propagating the culture, and can be accelerated by transferring plasmids between cells by conjugation or electroporation. Evolution can be further accelerated by use of mutator host cells or by seeding a culture of nonmutator host cells being evolved with mutator host cells and inducing intercellular transfer of plasmids by electroporation or conjugation. Preferably, mutator host cells used for seeding contain a negative selectable marker to facilitate isolation of a pure culture of the nonmutator cells being evolved. Selection/screening identifies cells bearing chromosomes and/or plasmids that have evolved toward acquisition of a desired function.

Subsequent rounds of recombination and selection/screening proceed in similar fashion to those described for plasmid-plasmid recombination. For example, further recombination can be effected by propagating cells surviving recombination in combination with electroporation or conjugative transfer of plasmids. Alternatively, plasmids bearing additional substrates for recombination can be introduced into the surviving cells. Preferably, such plasmids are from a different incompatibility group and bear a different selective marker than the original plasmids to allow selection for cells containing at least two different plasmids. As a further alternative, plasmid and/or chromosomal DNA can be isolated from a subpopulation of surviving cells and transfected into a second subpopulation. Chromosomal DNA can be cloned into a plasmid vector before transfection.

#### **8.6.1.2.2.5 VIRUS-CHROMOSOME RECOMBINATION**

As in the other methods described above, the virus is usually one that does not kill the cells, and is often a phage or phagemid. The procedure is substantially the same as for plasmid-chromosome recombination. Substrates for recombination are cloned into the vector. Vectors including the substrates can then be transfected into cells or in vitro



packaged and introduced into cells by infection. Viral genomes recombine with host chromosomes merely by propagating a culture. Evolution can be accelerated by allowing intercellular transfer of viral genomes by electroporation, or reinfection of cells by progeny virions. Screening/selection identifies cells having chromosomes and/or viral genomes that have evolved toward acquisition of a desired function.

There are several options for subsequent rounds of recombination. For example, viral genomes can be transferred between cells surviving selection/recombination by electroporation. Alternatively, viruses extruded from cells surviving selection/screening can be pooled and used to superinfect the cells at high multiplicity. Alternatively, fresh substrates for recombination can be introduced into the cells, either on plasmid or viral vectors.

#### **8.6.1.2.2.6 POOLWISE WHOLE GENOME RECOMBINATION**

Asexual evolution is a slow and inefficient process. Populations move as individuals rather than as a group. A diverse population is generated by mutagenesis of a single parent, resulting in a distribution of fit and unfit individuals. In the absence of a sexual cycle, each piece of genetic information for the surviving population remains in the individual mutants. Selection of the fittest results in many fit individuals being discarded, along with the genetically useful information they carry. Asexual evolution proceeds one genetic event at a time, and is thus limited by the intrinsic value of a single genetic event. Sexual evolution moves more quickly and efficiently. Mating within a population consolidates genetic information within the population and results in useful information being combined together.

The combining of useful genetic information results in progeny that are much more fit than their parents. Sexual evolution thus proceeds much faster by multiple genetic events. These differences are further illustrated herein. In contrast to sexual evolution, DNA stochastic &/or non-stochastic mutagenesis is the recursive mutagenesis, recombination, and selection of DNA sequences.

Sexual recombination in nature effects pairwise recombination and results in progeny that are genetic hybrids of two parents. In contrast, DNA stochastic &/or non-stochastic mutagenesis in vitro effects poolwise recombination, in which progeny are hybrids of multiple parental molecules. This is because DNA stochastic &/or non-stochastic mutagenesis effects many individual pairwise recombination events with each thermal cycle. After many cycles the result is a repetitively inbred population, with the "progeny" being the  $F_x$  ( for X cycles of stochastic &/or non-stochastic mutagenesis) of the original parental molecules. These progeny are potentially descendants of many or all of the original parents. One can graph to show a plot of the potential number of mutations an individual can accumulate by sequential, pairwise and poolwise recombination.

Poolwise recombination is an important feature to DNA stochastic &/or non-stochastic mutagenesis in that it provides a means of generating a greater proportion of the possible combinations of mutations from a single "breeding" experiment. In this way, the "genetic potential" of a population can be readily assessed by screening the progeny of a single DNA shuffling experiment.

For example, if a population consists of 10 single mutant parents, there are  $2^{10}=1024$  possible combinations of those mutations ranging from progeny having 0- 10 mutations. Of these 1024, only 56 will result from a single pairwise cross (i.e those having 0, 1, and 2 mutations). In nature the multiparent combinations will eventually arise after multiple random sexual matings, assuming no selection is imparted to remove some mutations from the population. In this way, sex effects the consolidation and sampling of all useful mutant combinations possible within a population. For the purposes of directed evolution, having the greatest number of mutant combinations entering a screen or selection is desirable so that the best progeny (i.e., according to the selection criteria used in the selection screen) is identified in the shortest possible time.

One challenge to in vivo and whole genome stochastic &/or non-stochastic mutagenesis is devising methods for effecting poolwise recombination or multiple repetitive pairwise recombination events. In crosses with a single pairwise cross per cycle

before screening, the ability to screen the "genetic potential" of the starting population is limited. For this reason, the rate of in vivo and whole genome stochastic &/or non-stochastic mutagenesis mediated cellular evolution would be facilitated by effecting poolwise recombination. Two strategies for poolwise recombination are described below (protoplast fusion and transduction).

#### 8.6.1.2.2.7 PROTOPLAST FUSION

Protoplast fusion (discussed supra) mediated whole genome stochastic &/or non-stochastic mutagenesis is one format that can directly effect poolwise recombination. Whole gene stochastic &/or non-stochastic mutagenesis is the recursive recombination of whole genomes, in the form of one or more nucleic acid molecule(s) (fragments, chromosomes, episomes, etc), from a population of organisms, resulting in the production of new organisms having distributed genetic information from at least two of the starting population of organisms. The process of protoplast fusion is further illustrated in herein.

Progeny resulting from the fusion of multiple parent protoplasts have been observed (Hopwood & Wright, 1978), however, these progeny are rare ( $10^{-4}$  -  $10^{-6}$ ). The low frequency is attributed to the distribution of fusants arising from two, three, four, etc parents and the likelihood of the multiple recombination events (6 crossovers for a four parent cross) that would have to occur for multiparent progeny to arise. Thus, it is useful to enrich for the multiparent progeny. This can be accomplished, e.g., by repetitive fusion or enrichment for multiply fused protoplasts. The process of poolwise fusion and recombination is further illustrated herein.

#### 8.6.1.2.2.8 REPETITIVE FUSION

Protoplasts of identified parental cells are prepared, fused and regenerated. Protoplasts of the regenerated progeny are then, without screening or enrichment, formed, fused and regenerated. This can be carried out for two, three, or more cycles before screening to increase the representation of multiparent progeny. The number of possible mutations/progeny doubles for each cycle. For example, if one cross produces predominantly progeny with 0, 1, and 2 mutations, a breeding of this population with itself

will produce progeny with 0, 1, 2, 3, and 4 mutations, the third cross up to eight, etc. The representation of the multiparent progeny from these subsequent crosses will not be as high as the single and double parent progeny, but it will be detectable and much higher than from a single cross. The repetitive fusion prior to screening is analogous to many sexual crosses within a population, and the individual thermal cycles of in vitro DNA stochastic &/or non-stochastic mutagenesis described supra. A factor effecting the value of this approach is the starting size of the parental population. As the population grows, it becomes more likely that a multiparent fusion will arise from repetitive fusions. For example, if 4 parents are fused twice, the 4 parent progeny will make up approximately 0.2% of the total progeny. This is sufficient to find in a population of 3000 (95% confidence), but better representation is preferable. If ten parents are fused twice >20% of the progeny will be four parent offspring.

#### **8.6.1.2.2.9 ENRICHMENT FOR MULTIPLE FUSED PROTOPLASTS**

After the fusion of a population of protoplasts, the fusants are typically diluted into hypotonic medium, to dilute out the fusing agent (e.g., 50% PEG). The fused cells can be grown for a short period to regenerate cell walls or separated directly and are then separated on the basis of size. This is carried out, e.g., by cell sorting, using light dispersion as an estimate of size, to isolate the largest fusants. Alternatively the fusants can be sorted by FACS on the basis of DNA content. The large fusants or those containing more DNA result from the fusion of multiple parents and are more likely to segregate to multiparent progeny. The enriched fusants are regenerated and screened directly or the progeny are fused recursively as above to further enrich the population for diverse mutant combinations.

#### **8.6.1.2.2.10 TRANSDUCTION**

Transduction can theoretically effect poolwise recombination, if the transducing phage particles contain predominantly host genomic DNA rather than phage DNA. If phage DNA is overly represented, then most cells will receive at least one undesired phage genome.

Phage particles generated from locked-in-prophage (supra) are useful for this purpose. A population of cells is infected with an appropriate transducing phage, and the lysate is collected and used to infect the same starting population. A high multiplicity of infection is employed to deliver multiple genomic fragments to each infected cell, thereby increasing the chance of producing recombinants containing mutations from more than two parent genomes.

The resulting transductants are recovered under conditions where phage can not propagate e.g., in the presence of citrate. This population is then screened directly or infected again with phage, with the resulting transducing particles being used to transduce the first progeny. This would mimic recursive protoplast fusion, multiple sexual recombination, and in vitro DNA stochastic &/or non-stochastic mutagenesis.

#### **8.6.1.2.2.11 METHODS FOR WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS BY BLIND FAMILY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF PARSED GENOMES AND RECURSIVE CYCLES OF FORCED INTEGRATION AND EXCISION BY HOMOLOGOUS RECOMBINATION, AND SCREENING FOR IMPROVED PHENOTYPES**

In vitro methods have been developed to reassemble single genes and operons, as set forth, e.g., herein. "Family" stochastic &/or non-stochastic mutagenesis of homologous genes within species and from different species is also an effective methods for accelerating molecular evolution. This section describes additional methods for extending these methods such that they can be applied to whole genomes.

In some cases, the genes that encode rate limiting steps in a biochemical process, or that contribute to a phenotype of interest are known. This method can be used to target family stochastic &/or non-stochastic mutagenized libraries to such loci, generating libraries of organisms with high quality family stochastic &/or non-stochastic mutagenized libraries of alleles at the locus of interest. An example of such a gene would be the evolution of a host chaperonin to more efficiently chaperone the folding of an overexpressed protein in *E. coli*.

The goals of this process are to reassemble homologous genes from two or more species and to then integrate the stochastic &/or non-stochastic mutagenized genes into the chromosome of a target organism.

Integration of multiple stochastic &/or non-stochastic mutagenized genes at multiple loci can be achieved using recursive cycles of integration (generating duplications), excision (leaving the improved allele in the chromosome) and transfer of additional evolved genes by serially applying the same procedure.

In the first step, genes to be stochastic &/or non-stochastic mutagenized into suitable bacterial vectors are subcloned. These vectors can be plasmids, cosmids, BACS or the like. Thus, fragments from 100 bp to 100 kb can be handled. Homologous fragments are then "family stochastic &/or non-stochastic mutagenized" together (i.e. homologous fragments from different species or chromosomal locations are homologously recombined). As a simple case, homologs from two species (say, *E. coli* and *Salmonella*) are cloned, family stochastic &/or non-stochastic mutagenized in vitro and cloned into an allele replacement vector (e.g., a vector with a positively selectable marker, a negatively selectable marker and conditionally active origin of replication). The basic strategy for whole genome family stochastic &/or non-stochastic mutagenesis of parsed (subcloned) genomes is additionally set forth herein.

The vectors are transfected into *E. coli* and selected, e.g., for drug resistance. Most drug resistant cells should arise by homologous recombination between a family stochastic &/or non-stochastic mutagenized insert and a chromosomal copy of the cloned insert. Colonies with improved phenotype are screened (e.g., by mass spectroscopy for enzyme activity or small molecule production, or a chromogenic screen, or the like, depending on the phenotype to be assayed). Negative selection (i.e. sue selection) is imposed to force excision of tandem duplication. Roughly half C, of the colonies should retain the improved phenotype. Importantly, this process regenerates a "clean" chromosome in which the wild type locus is replaced with a family stochastic &/or non-stochastic mutagenized fragment

that encodes a beneficial allele. Since the chromosome is "clean" (i.e., has no vector sequences), other improved alleles can also be moved into this point on the chromosome by homologous recombination.

Selection or screening for improved phenotype can occur either after step 3 or step 4. If selection or screening takes place after step 3, then the improved allele can be conveniently moved to other strains by, for example, P I transduction. One can then regenerate a strain containing the improved allele but lacking vector sequences by "negative selection" against the suc marker. In subsequent rounds, independently identified improved variants of the gene can be sequentially moved into the improved strain (e.g., by P I transduction of the drug marked tandem duplication above). Transductants are screened for further improvement in phenotype by virtue of receiving the transduced tandem duplication, which itself contains the family stochastic &/or non-stochastic mutagenized genetic material. Negative selection is again imposed and the process of stochastic &/or non-stochastic mutagenesis the improved strain is recursively repeated as desired.

Although this process was described with reference to targeting a gene or genes of interest, it can be used "blindly," making no assumptions about which locus is to be targeted. This procedure is set forth herein. For example, the whole genome of an organism of interest is cloned into manageable fragments (e.g., 10 kb for plasmid-based methods). Homologous fragments are then isolated from related species. Forced recombination with chromosomal homologs creates chimeras.

#### **8.6.1.2.2.12 METHODS FOR HIGH THROUGHPUT FAMILY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF GENES**

For *E. coli*., cloning the genome in 10 kb fragments requires about 300 clones. The homologous fragments are isolated, e.g., from *Salmonella*. This gives roughly three hundred pairs of homologous fragments. Each pair is family stochastic &/or non-stochastic mutagenized and the stochastic &/or non-stochastic mutagenized fragments are cloned into an allele replacement vector. The inserts are integrated into the *E. coli* genome

as described above. A global screen is made to identify variants with an improved phenotype. This serves as the basis collection of improvements that are to be stochastic &/or non-stochastic mutagenized to produce a desired strain. The stochastic &/or non-stochastic mutagenesis of these independently identified variants into one super strain is done as described above.

Family stochastic &/or non-stochastic mutagenesis has been shown to be an efficient method for creating high quality libraries of genetic variants. Given a cloned gene from one species, it is of interest to quickly and rapidly isolate homologs from other species, and this process can be rate limiting. For example, if one wants to perform family stochastic &/or non-stochastic mutagenesis on an entire genome, one may need to construct hundreds to thousands of individual family stochastic &/or non-stochastic mutagenized libraries.

In this embodiment, a gene of interest is optionally cloned into a vector in which ssDNA can be made. An example of such a vector is a phagemid vector with an M13 origin of replication. Genomic DNA or cDNA from a species of interest is isolated, denatured, annealed to the phagemid, and then enzymatically manipulated to clone it. The cloned DNA is then used to family reassemble with the original gene of interest. PCR based formats are also available. These formats require no intermediate cloning steps, and are, therefore, of particular interest for high throughput applications.

Alternatively, the gene of interest can be fished out using purified RecA protein. The gene of interest is PCR amplified using primers that are tagged with an affinity tag such as biotin, denatured, then coated with RecA protein (or an improved variant thereof). The coated ssDNA is then mixed with a gDNA plasmid library. Under the appropriate conditions, such as in the presence of non-hydrolyzable rATP analogs, RecA will catalyze the hybridization of the RecA coated gene (ssDNA) in the plasmid library. The heteroduplex is then affinity purified from the non-hybridizing plasmids of the gene library by adsorption of the labeled PCR products and its associated homologous DNA to an appropriate affinity matrix.



The homologous DNA is used in a family stochastic &/or non-stochastic mutagenesis reaction for improvement of the desired function. Stochastic &/or non-stochastic mutagenesis the *E. coli* chaperonin gene DnaJ with other homologs is described below as an example. The example can be generalized to any other gene, including eukaryotic genes such as plant or animal genes (including mammalian genes), by following the format described.

As a first step, the *E. coli* DnaJ gene is cloned into an M13 phagemid vector. ssDNA is then produced, preferably in a dut(-) ung(-) strain so that Kunkel site directed mutagenesis protocols can be applied. Genomic DNA is then isolated from a non-*E. coli* source, such as *Salmonella* and *Yersinia Pestis*. The bacterial genomic DNAs are denatured and reannealed to the phagemid ssDNA (e.g., about 1 microgram of ssDNA). The reannealed product is treated with an enzyme such as Mung Bean nuclease that degrades ssDNA as an exonuclease but not as an endonuclease (the nuclease does not degrade mismatched DNA that is embedded in a larger annealed fragment). The standard Kunkel site directed mutagenesis protocol is used to extend the fragment and the target cells are transformed with the resulting mutagenized DNA.

In a first variation on the above, the procedure is adapted to the situation where the target gene or genes of interest are unknown. In this variation, the whole genome of the organism of interest is cloned in fragments (e.g., of about 10 kb each) into a phagemid. Single stranded phagemid DNA is then produced. Genomic DNA from the related species is denatured and annealed to the phagemids. Mung bean nuclease is used to trim away unhybridized DNA ends. Polymerase plus ligase is used to fill in the resulting gapped circles.

These clones are transformed into a mismatch repair deficient strain. When the mismatched molecules are replicated in the bacteria, most colonies contain both the *E. coli* and the homologous fragment. The two homologous genes are then isolated from the

colonies (e.g., either by standard plasmid purification or colony PCR) and stochastic &/or non-stochastic mutagenized.

Another approach to generating chimeras that requires no in vitro stochastic &/or non-stochastic mutagenesis is simply to clone the Salmonella genome into an allele replacement vector, transform *E. coli*, and select for chromosomal integrants. Homologous recombination between Salmonella genes and *E. coli* homologs generate stochastic &/or non-stochastic mutagenized chimeras. A global screen is done to screen for improved phenotypes. Alternately, recursive transformation and recombination is performed to increase diversity prior to screening. If colonies with improved phenotypes are obtained, it is verified that the improvement is due to allele replacement by P I transduction into a fresh strain and counterscreening for improved phenotype. A collection of such improved alleles can then be combined into one strain using the methods for whole genome stochastic &/or non-stochastic mutagenesis by blind family stochastic &/or non-stochastic mutagenesis of parsed genomes as set forth herein. Additionally, once these loci are identified, it is likely that further rounds of stochastic &/or non-stochastic mutagenesis and screening will yield further improvements. This could be done by cloning the chimeric gene and then using the methods described in this disclosure to breed the gene with homologs from many different strains of bacteria.

In general, the transformants contain clones of the homologue of the target gene (e.g., *E. coli* DnaJ in the example above). Mismatch repair in vivo results in a decrease in diversity of the gene. There are at least two solutions to this. First, transduction can be performed into a mismatch repair deficient strain. Alternatively or in addition, the M13 template DNA can be selectively degraded, leaving the cloned homologue. This can be done using methods similar to the standard Eckstein site directed mutagenesis technique (General texts which describe general molecular biological techniques useful herein, including mutagenesis, include Sambrook et al., *Molecular Cloning - A Laboratory Manual* (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and *Current Protocols in Molecular Biology*, F.M. Ausubel et

al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel").

This method relies on incorporation of alpha thiol modified dNTPs during synthesis of the new strand followed by selective degradation of the template and resynthesis of the template strand. In one embodiment, the template strand is grown in a *dut(-) ung(-)* strain so that uracil is incorporated into the phagemid DNA. After extension as noted above (and before transformation) the DNA is treated with uracil glycosylate and an apurinic site endonuclease such as Endo III or Endo IV. The treated DNA is then treated with a processive exonuclease that resects from the resulting gaps while leaving the other strand intact (as in Eckstein mutagenesis). The DNA is polymerized and ligated. Target cells are then transformed. This process enriches for clones encoding the homologue which is not derived from the target (i.e., in the example above, the non- *E. coli* homologue).

An analogous procedure is optionally performed in a PCR format. As applied to the DnaJ illustration above, DnaJ DNA is amplified by PCR with primers that build 30-mer priming sites on each end. The PCR is denatured and annealed with an excess of *Salmonella* genomic DNA. The *Salmonella* DnaJ gene hybridizes with the *E. coli* homologue. After treatment with Mung Bean nuclease, the resulting mismatched hybrid is PCR amplified with the flanking 30-mer primers. This PCR product can be used directly for family stochastic &/or non-stochastic mutagenesis. As genomics provides an increasing amount of sequence information, it is increasingly possible to directly PCR amplify homologs with designed primers. For example, given the sequence of the *E. coli* genome and of a related genome (i.e. *Salmonella*), each genome can be PCR amplified with designed primers in, e.g., 5 kb fragments. The homologous fragments can be put together in a pairwise fashion for stochastic &/or non-stochastic mutagenesis. For genome stochastic &/or non-stochastic mutagenesis, the stochastic &/or non-stochastic mutagenized products are cloned into the allele replacement vector and bred into the genome as described supra.

#### 8.6.1.2.2.13 HYPER-RECOMBINOGENIC RecA CLONES

The invention further provides hyper-recombinogenic RecA proteins (see, the examples below). It is fully expected that one of skill can make a variety of related recombinogenic proteins given the disclosed sequences.

Standard molecular biological techniques can be used to make nucleic acids which comprise the given nucleic acids, e.g., by cloning the nucleic acids into any known vector. Examples of appropriate cloning and sequencing techniques, and instructions sufficient to direct persons of skill through many cloning exercises are found in Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al. (1989) Molecular Cloning - A Laboratory Manual (2nd ed.) Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor Press, NY, (Sambrook); and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (1994 Supplement) (Ausubel). Product information from manufacturers of biological reagents and experimental equipment also provide information useful in known biological methods. Such manufacturers include the SIGMA chemical company (Saint Louis, MO); R&D systems (Minneapolis, MN), Pharmacia LKB Biotechnology (Piscataway, NJ), CLONTECH Laboratories, Inc. (Palo Alto, CA), Chem Genes Corp., Aldrich Chemical Company (Milwaukee, WI), Glen Research, Inc., GIBCO BRL Life Technologies, Inc. (Gaithersburg, MI), Fluka Chemica-Biochemika Analytika (Fluka Chemie AG, Buchs, Switzerland), Invitrogen, San Diego, CA, and Applied Biosystems (Foster City, CA), as well as many other commercial sources known to one of skill.

It will be appreciated that conservative substitutions of the given sequences can be used to produce nucleic acids which encode hyperrecombinogenic clones. "Conservatively modified variations" of a particular nucleic acid sequence refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic

acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon, for methionine) can be modified to yield a functionally identical molecule by standard techniques. Accordingly, each "silent variation" of a nucleic acid which encodes a polypeptide is implicit in any described sequence. Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following six groups each contain amino acids that are conservative substitutions for one another: 1) Alanine (A), Serine (S), Threonine (T); 2) Aspartic acid (D), Glutamic acid (E); 3) Asparagine (N), Glutamine (Q); 4) Arginine (R), Lysine (K); 5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); and 6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W). See also, Creighton (1984) *Proteins* W.H. Freeman and Company. Finally, the addition of sequences which do not alter the encoded activity of a nucleic acid molecule, such as a non-functional sequence is a conservative modification of the basic nucleic acid.

One of skill will appreciate that many conservative variations of the nucleic acid constructs disclosed yield a functionally identical construct. For example, due to the degeneracy of the genetic code, "silent substitutions" (ie., substitutions of a nucleic acid sequence which do not result in an alteration in an encoded polypeptide) are an implied feature of every nucleic acid sequence which encodes an amino acid. Similarly, "conservative amino acid substitutions," in one or a few amino acids in an amino acid sequence of a packaging or packageable construct are substituted with different amino

acids with highly similar properties, are also readily identified as being highly similar to a disclosed construct. Such conservatively substituted variations of each explicitly disclosed sequence are a feature of the present invention.

Nucleic acids which hybridize under stringent conditions to the nucleic acids in the figures are a feature of the invention. "Stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) Laboratory Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic Acid Probes part I chapter 2 "overview of principles of hybridization and the strategy of nucleic acid probe assays", Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5C lower than the thermal melting point (T<sub>m</sub>) for the specific sequence at a defined ionic strength and pH. The T<sub>m</sub> is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T<sub>m</sub> for a particular probe. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

Nucleic acids which do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

Finally, preferred nucleic acids encode hyper-recombinogenic RecA proteins which are at least one order of magnitude (10 times) as active as a wild-type RecA protein in a standard assay for Rec A activity.

#### **8.6.1.2.2.14 RecE/RecT MEDIATED STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS IN VIVO**

Like recA, recE and recT (or their homologues, for example the lambda recombination proteins  $\text{red}^+$  and  $\text{red}^-$ ) can stimulate homologous recombination in vivo. See, Muyrers et al. (1999) *Nucleic Acids Res* 27(6):1555-7 and Zhang et al. (1998) *Nat Genet* (2):123-8 Hyper-recombinogenic recE and recT are evolved by the same method as described for recA. Alternatively, variants with increased recombinogenicity are selected by their ability to cause recombination between a suicide vector (lacking an origin of replication) carrying a selectable marker, and a homologous region in either the chromosome or a stably- maintained episome.

A plasmid containing recA and recE genes is stochastic &/or non-stochastic mutagenized (either using these genes as single starting points, or by family stochastic &/or non-stochastic mutagenesis (with for example  $\text{red}^+$  and  $\text{red}^-$ , or other homologous genes identified from available sequence databases). This stochastic &/or non-stochastic mutagenized library is then cloned into a vector with a selectable marker and transformed into an appropriate recombination-deficient strain. The library of cells would then be transformed with a second selectable marker, either borne on a suicide vector or as a linear DNA fragment with regions at its ends that are homologous to a target sequence (either in the plasmid or in the host chromosome). Integration of this marker by homologous recombination is a selectable event, dependent on the activity of the recE and recT gene products. The recE / recT genes are isolated from cells in which homologous recombination has occurred. The process is repeated several times to enrich for the most efficient variants before the next round of stochastic &/or non-stochastic mutagenesis is performed. In addition, cycles of recombination without selection can be performed to increase the diversity of a cell population prior to selection.

Once hyper-recombinogenic recE / recT genes are isolated they are used as described for hyper-recombinogenic recA. For example they are expressed (constitutively or conditionally) in a host cell to facilitate homologous recombination between variant gene fragments and homologues within the host cell. They are alternatively introduced by

microinjection, biolistics, lipofection or other means into a host cell at the same time as the variant genes.

Hyper-recombinogenic recE/ recT (either of bacterial / phage origin, or from plant homologues) are useful for facilitating homologous recombination in plants. They are, for example, cloned into the Agrobacterium cloning vector, where they are expressed upon entry into the plant, thereby stimulating homologous recombination in the recipient cell.

In a preferred embodiment, recE/ recT are used and or generated in mutS strains.

#### 8.6.1.2.2.15 MULTI-CYCLIC RECOMBINATION

As noted, protoplast fusion is an efficient means of recombining two microbial genomes. The process reproducibly results in about 10% of a non-selected population being recombinant chimeric organisms.

Protoplasts are cells that have been stripped of their cell walls by treatment in hypotonic medium with cell wall degrading enzymes. Protoplast fusion is the induced fusion of the membranes of two or more of these protoplasts by fusogenic agents such as polyethylene glycol. Fusion results in cytoplasmic mixing and places the genomes of the fused cells within the same membrane. Under these conditions recombination between the genomes is frequent.

The fused protoplasts are regenerated, and, during cell division, single genomes segregate into each daughter cell. Typically, 10% of these daughter cells have genomes that originate partially from more than one of the original parental protoplast genomes.

This result is similar to that of the crossing over of sister chromatids in eukaryotic cells during prophase of meiosis II. The percentage of daughter cells that are recombinant is just lower after protoplast fusion. While protoplast fusion does result in efficient recombination, the recombination predominantly occurs between two cells as in sexual recombination.



In order to efficiently generate libraries of whole genome stochastic &/or non-stochastic mutagenized libraries, daughter cells having genetic information originating from multiple parents are made.

In vitro DNA stochastic &/or non-stochastic mutagenesis results in the efficient poolwise recombination of multiple homologous DNA sequences. The stochastic &/or non-stochastic mutagenesis of full length genes from a mixed pool of small gene fragments requires multiple annealing and elongation cycles, the thermal cycles of the primerless PCR reaction. During each thermal cycle, many pairs of fragments anneal and are extended to form a combinatorial population of larger chimeric DNA fragments. After the first cycle of stochastic &/or non-stochastic mutagenesis, chimeric fragments contain sequences originating from two different parent genes. This is similar to the result of a single sexual cycle within a population, pairwise cross, or protoplast fusion. During the second cycle, these chimeric fragments can anneal with each other, or with other small fragments, resulting in chimeras originating from up to four different parental sequences.

This second cycle is analogous to the entire progeny from a single sexual cross inbreeding with itself. Further cycles will result in chimeras originating from 8, 16, 32, etc parental sequences and are analogous to further inbreedings of the progeny population. The power of in vitro DNA stochastic &/or non-stochastic mutagenesis is that a large combinatorial library can be generated from a single pool of DNA fragments stochastic &/or non-stochastic mutagenized by these recursive pairwise "matings." As described above, in vivo stochastic &/or non-stochastic mutagenesis strategies, such as protoplast fusion, result in a single pairwise mating reaction. Thus, to generate the level of diversity obtained by in vitro methods, in vivo methods are carried out recursively. That is, a pool of organisms is recombined and the progeny pooled, without selection, and then recombined again. This process is repeated for sufficient cycles to result in progeny having multiple parental sequences.

Described below is a method used to reassemble four strains of *Streptomyces coelicolor*. From the initial four strains each containing a unique nutritional marker, three to four rounds of recursive pooled protoplast fusion was sufficient to generate a population of stochastic &/or non-stochastic mutagenized organisms containing all 16 possible combinations of the four markers. This represents a  $10^6$  fold improvement in the generation of four parent progeny as compared to a single pooled fusion of the four strains.

Protoplasts were generated from several strains of *S. coelicolor*, pooled and fused. Mycelia were regenerated and allowed to sporulate. The spores were collected, allowed to grow into Mycelia, formed into protoplasts, pooled and fused and the process repeated for three to four rounds. the resulting spores were then subject to screening.

The basic protocol for generating a whole genome stochastic &/or non-stochastic mutagenized library from four *S. coelicolor* strains, each having one of four distinct markers, was as follows. Four mycelial cultures, each of a strain having one of four different markers, were grown to early stationary phase. The mycelia from each were harvested by centrifugation and washed. Protoplasts from each culture were prepared as follows. Approximately  $10^9$  *S. coelicolor* spores were inoculated into 50ml YEME with 0.5% Glycine in a 250ml baffled flask. The spores were incubated at 30C for 36-40 hours in an orbital shaker. Mycelium were verified using a microscope. Some strains needed an additional day of growth. The culture was transferred into a 50ml tube and centrifuged at 4,000 rpm for 10 min. The mycelium were twice washed with 10.3% sucrose and centrifuged at 4,000 rpm for 10 min. (mycelium can be stored at about 80C after wash). 5ml of lysozyme was added to the about 0.5g of mycelium pellet. The pellet was suspended and incubated at 30C for 20-60 min., with gentle shaking every 10 min. The microscope was checked for protoplasting every 20 min. Once the majority were protoplasts, protoplasting was stopped by adding 10ml of P buffer. The protoplasts were filtered through cotton and the protoplast spun down at 3,000rpm for 7 min at room temperature. The supernatant was discarded and the protoplast gently resuspended, adding a suitable amount of P buffer according to the pellet size (usually about 500W). Ten-fold

serial dilutions were made in P buffer, and the protoplasts counted at a  $10^{-2}$  dilution. Protoplasts were adjusted to  $10^{10}$  protoplasts per ml.

The protoplasts from each culture were quantitated by microscopy.  $10^8$  protoplast from each culture were mixed in the same tube, washed, and then fused by the addition of 50% PEG. The fused protoplasts were diluted and plated regeneration medium and incubated until the colonies were sporulating (four days). Spores were harvested and washed. These spores represent a pool of all the recombinants and parents from the fusion.

A sample of the pooled spores was then used to inoculate a single liquid culture. The culture was grown to early stationary phase, the mycelia harvested, and protoplasts prepared.  $10^8$  protoplasts from this "mycelial library" were then fused with themselves by the addition of 50%PEG. The protoplast fusion/regeneration/harvesting/protoplast preparation steps were repeated two times. The spores resulting from the fourth round of fusion were considered the "whole genome stochastic &/or non-stochastic mutagenized library" and they were screened for the frequency of the 16 possible combinations of the four markers

In particular, adding rounds of recombination prior to selection produced significant increases in the number of clones which incorporated all four of the relevant selectable markers, indicating that the population became increasingly diverse by recursive pooling and sporulation.

The four strains of the four parent stochastic &/or non-stochastic mutagenesis were each auxotrophic for three and prototrophic for one of four possible nutritional markers: arginine (A), cystine (C), proline (P), and/or uracil (U). Spores from each fusion were plated in each of the 16 possible combinations of these four nutrients, and the percent of the population growing on a particulate medium was calculated as the ratio of those colonies from a selective plate to those growing on a plate having all four nutrients (all variants grow on the medium having all four nutrients, thus the colonies from this plate thus represent the total viable population). The corrected percentages for each of the no,

one, two, and three marker phenotypes were determined by subtracting the percentage of cells having additional markers that might grow on the medium having "unnecessary" nutrients. For example, the number of colonies growing on no additional nutrients (the prototroph) was subtracted from the number of colonies growing on any plate requiring nutrients.

#### **8.6.1.2.2.16 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS THROUGH ORGANIZED HETERODUPLEX STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

A new procedure to optimize phenotypes of interests by heteroduplex stochastic &/or non-stochastic mutagenesis of cosmid libraries of the organism of choice, is provided. This procedure does not require protoplast fusion and is applicable to bacteria for which well-established genetic systems are available, including cosmid cloning, transformation, in vitro packaging/transfection and plasmid transfer/mobilization. Microorganism that can be improved by these methods include *Escherichia coli*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Pseudomonas* spp., *Rhizobium* spp., *Xanthomonas* spp., and other gram-negative organisms. This method is also applicable to Gram-positive microorganisms.

In step A, Chromosomal DNA of the organism to be improved is digested with suitable restriction enzymes and ligated into a cosmid. The cosmid used for cosmid-based heteroduplex guided whole genome stochastic &/or non-stochastic mutagenesis has at least two rare restriction enzyme recognition sites (e.g. *Sfr* and *NotI*) to be used for linearization in subsequent steps. Sufficient cosmids to represent the complete chromosome are purified and stored in 96-well microtiter dishes. In step B, small samples of the library are mutagenized in vitro using hydroxylamine or other mutagenic chemicals. In step C, a sample from each well of the mutagenized collection is used to transfect the target cells. In step D, the transfectants are assayed (as a pool from each mutagenized sample-well) for phenotypic improvements. Positives from this assay indicate that a cosmid from a particular well can confer phenotypic improvements and thus contain large genomic fragments that are suitable targets for heteroduplex mediated stochastic &/or non-

stochastic mutagenesis. In step E, the transfected cells harboring a mutant library of the identified cosmid(s) are separated by plating on solid media and screened for independent mutants conferring an improved phenotype. In step F, DNA from positive cells is isolated and pooled by origin. In step G, the selected cosmid pools are divided so that one sample can be digested with Sfr and the other with NotI. These samples are pooled, denatured, reannealed, and religated.

In step K target cells are transfected with the resulting heteroduplexes and propagated to allow "recombination" to occur between the strands of the heteroduplexes in vivo. The transfectants can be screened (the population will represent the pairwise recombinants) or, commonly, as represented by step 1, the recombined cosmids are further stochastic &/or non-stochastic mutagenized by recursive in vitro heteroduplex formation and in vivo recombination (to generate a complete combinatorial library of the possible mutations) prior to screening. An additional mutagenesis step could also be added for increased diversity during the stochastic &/or non-stochastic mutagenesis process.

In step J, once several cosmids harboring different distributed loci have been improved, they are combined into the same host by chromosome integration. This organism can be used directly or subjected to a new round of heteroduplex guided whole genome stochastic &/or non-stochastic mutagenesis.

## **8.7. SPECIALIZED METHODS**

### **8.7.1 TARGETED STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS-HOT SPOTS**

In one aspect, targeted homologous genes are cloned into specific regions of the genome (e.g., by homologous recombination or other targeting procedures) which are known to be recombination "hot spots" (i.e., regions showing elevated levels of recombination compared to the average level of recombination observed across an entire genome), or known to be proximal to such hot spots. The resulting recombinant strains are mated recursively. During meiotic recombination, homologous recombinant genes

recombine, thereby increasing the diversity of the genes. After several cycles of recombination by recursive mating, the resulting cells are screened.

### 8.7.2 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS USING YEASTS

Yeasts are subspecies of fungi that grow as single cells. Yeasts are used for the production of fermented beverages and leavening, for production of ethanol as a fuel, low molecular weight compounds, and for the heterologous production of proteins and enzymes (see accompanying list of yeast strains and their uses). Commonly used strains of yeast include *Saccharomyces cerevisiae*, *Pichia* sp., *Candida* sp. and *Schizosaccharomyces pombe*.

Several types of vectors are available for cloning in yeast including integrative plasmid (YIp), yeast replicating plasmid (YRp, such as the 2 circle based vectors), yeast episomal plasmid (YEp), yeast centromeric plasmid (YCp), or yeast artificial chromosome (YAC). Each vector can carry markers useful to select for the presence of the plasmid such as LUE2, URA3, and HIS3, or the absence of the plasmid such as URA3 (a gene that is toxic to cells grown in the presence of 5-fluoro orotic acid).

Many yeasts have a sexual cycle and asexual (vegetative) cycles. The sexual cycle involves the recombination of the whole genome of the organism each time the cell passes through meiosis. For example, when diploid cells of *S. cerevisiae* are exposed to nitrogen and carbon limiting conditions, diploid cells undergo meiosis to form asci. Each ascus holds four haploid spores, two of mating type "a" and two of mating type "α". Upon return to rich medium, haploid spores of opposite mating type mate to form diploid cells once again. Asiospores of opposite mating type can mate within the ascus, or if the ascus is degraded, for example with zymolase, the haploid cells are liberated and can mate with spores from other asci. This sexual cycle provides a format to reassemble endogenous genomes of yeast and/or exogenous fragment libraries inserted into yeast vectors. This process results in swapping or accumulation of hybrid genes, and for the stochastic &/or non-stochastic mutagenesis of homologous sequences shared by mating cells.

Yeast strains having mutations in several known genes have properties useful for stochastic &/or non-stochastic mutagenesis. These properties include increasing the frequency of recombination and increasing the frequency of spontaneous mutations within a cell. These properties can be the result of mutation of a coding sequence or altered expression (usually overexpression) of a wildtype coding sequence. The HO nuclease effects the transposition of HMLa/ and HMRA/ to the MAT locus resulting in mating type switching. Mutants in the gene encoding this enzyme do not switch their mating type and can be employed to force crossing between strains of defined genotype, such as ones that harbor a library or have a desired phenotype and to prevent in breeding of starter strains. PMS1, MLH1, MSH2, MSH6 are involved in mismatch repair. Mutations in these genes all have a mutator phenotype (Chambers et al., Mol. Cell. Biol. 16, 6110-6120 (1996)). Mutations in TOP3 DNA topoisomerase have a 6-fold enhancement of interchromosomal homologous recombination (Bailis et al., Molecular and Cellular Biology 12, 4988-4993 (1992)). The RAD50-57 genes confer resistance to radiation. Rad3 functions in excision of pyrimidine dimers. RAD52 functions in gene conversion. RAD50, MRE11, XRS2 function in both homologous recombination and illegitimate recombination. HOP1, RED1 function in early meiotic recombination (Mao-Draayer, Genetics 144, 71-86) Mutations in either HOP1 or RED 1 reduce double stranded breaks at the HIS2 recombination hotspot. Strains deficient in these genes are useful for maintaining stability in hyper recombinogenic constructs such as tandem expression libraries carried on YACs. Mutations in HPR1 are hyperrecombinogenic. HDF1 has DNA end binding activity and is involved in double stranded break repair and V(D)J recombination.

Strains bearing this mutation are useful for transformation with random genomic fragments by either protoplast fusion or electroporation. Kar-1 is a dominant mutation that prevents karyogamy. Kar-1 mutants are useful for the directed transfer of single chromosomes from a donor to a recipient strain. This technique has been widely used in the transfer of YACs between strains, and is also useful in the transfer of evolved genes/chromosomes to other organisms (Markie, YAC Protocols, (Humana Press,

Totowa, NJ, 1996). HOT1 is an *S. cerevisiae* recombination hotspot within the promoter and enhancer region of the rDNA repeat sequences. This locus induces mitotic recombination at adjacent sequences- presumably due to its high level transcription. Genes and/or pathways inserted under the transcriptional control of this region undergo increased mitotic recombination. The regions surrounding the *arg 4* and *his 4* genes are also recombination hot spots, and genes cloned in these regions have an increased probability of undergoing recombination during meiosis.

Homologous genes can be cloned in these regions and stochastic &/or non-stochastic mutagenized in vivo by recursively mating the recombinant strains. CDC2 encodes polymerase and is necessary for mitotic gene conversion. Overexpression of this gene can be used in a reassembler or mutator strain. A temperature sensitive mutation in CDC4 halts the cell cycle at G1 at the restrictive temperature and could be used to synchronize protoplasts for optimized fusion and subsequent recombination.

As with filamentous fungi, the general goals of stochastic &/or non-stochastic mutagenesis yeast include improvement in yeast as a host organism for genetic manipulation, and as a production apparatus for various compounds. One desired property in either case is to improve the capacity of yeast to express and secrete a heterologous protein. The following example describes the use of stochastic &/or non-stochastic mutagenesis to evolve yeast to express and secrete increased amounts of RNase A.

RNase A catalyzes the cleavage of the P-O<sub>5'</sub> bond of RNA specifically after pyrimidine nucleotides. The enzyme is a basic 124 amino acid polypeptide that has 8 half cystine residues, each required for catalysis. YEpWL-RNase A is a vector that effects the expression and secretion of RNaseA from the yeast *S. cerevisiae*, and yeast harboring this vector secrete 1-2 mg of recombinant RNase A per liter of culture medium (del Cardayre et al., Protein Engineering 8(3):26, 1-273 (1995)). This overall yield is poor for a protein heterologously expressed in yeast and can be improved at least 10-100 fold by stochastic &/or non-stochastic mutagenesis. The expression of RNaseA is easily detected by several plate and microtitre plate assays (del Cardayre & Raines, Biochemistry 33, 6031-6037



1994)). Each of the described formats for whole genome stochastic &/or non-stochastic mutagenesis can be used to reassemble a strain of *S. cerevisiae* harboring YepWL-RNase A, and the resulting cells can be screened for the increased secretion of RNase A into the medium. The new strains are cycled recursively through the stochastic &/or non-stochastic mutagenesis format, until sufficiently high levels of RNase A secretion is observed. The use of RNase A is particularly useful since it not only requires proper folding and disulfide bond formation but also proper glycosylation. Thus numerous components of the expression, folding, and secretion systems can be optimized. The resulting strain is also evolved for improved secretion of other heterologous proteins.

### **8.7.3 REASSEMBLE TO INCREASE TOLERANCE OF YEAST TO ETHANOL**

Another goal of stochastic &/or non-stochastic mutagenesis yeast is to increase the tolerance of yeast to ethanol. Such is useful both for the commercial production of ethanol, and for the production of more alcoholic beers and wines. The yeast strain to be stochastic &/or non-stochastic mutagenized acquires genetic material by exchange or transformation with other strain(s) of yeast, which may or may not be known to have superior resistance to ethanol. The strain to be evolved is stochastic &/or non-stochastic mutagenized and shufflants are selected for capacity to survive exposure to ethanol. Increasing concentrations of ethanol can be used in successive rounds of stochastic &/or non-stochastic mutagenesis. The same principles can be used to reassemble baking yeasts for improved osmotolerance.

### **8.7.4 CAPACITY TO GROW UNDER DESIRED NUTRITIONAL CONDITIONS**

Another desired property of stochastic &/or non-stochastic mutagenesis yeast is capacity to grow under desired nutritional conditions. For example, it is useful to yeast to grow on cheap carbon sources such as methanol, starch, molasses, cellulose, cellobiose, or xylose depending on availability. The principles of stochastic &/or non-stochastic mutagenesis and selection are similar to those discussed for filamentous fungi.

### 8.7.5 TO PRODUCE SECONDARY METABOLITES

Another desired property is capacity to produce secondary metabolites naturally produced by filamentous fungi or bacteria. Examples of such secondary metabolites are cyclosporin A, taxol, and cephalosporins. The yeast to be evolved undergoes genetic exchange or is transformed with DNA from organism(s) that produce the secondary metabolite. For example, fungi producing taxol include *Taxomyces andreanae* and *Pestalotopsis microspora* (Stierle et al., Science 260, 214-216 (1993); Strobel et al., Microbiol. 142, 435440 (1996)). DNA can also be obtained from trees that naturally produce taxol, such as *Taxus brevifolia*. DNA encoding one enzyme in the taxol pathway, taxadiene synthase, which it is believed catalyzes the committed step in taxol biosynthesis and may be rate limiting in overall taxol production, has been cloned (Wildung & Croteau, J Biol Chem. 271, 9201-4 (1996)). The DNA is then stochastic &/or non-stochastic mutagenized, and shufflants are screened/selected for production of the secondary metabolite. For example, taxol production can be monitored using antibodies to taxol, by mass spectroscopy or UV spectrophotometry. Alternatively, production of intermediates in taxol synthesis or enzymes in the taxol synthetic pathway can be monitored. Concetti & Ripani, Biol Chem. Hoppe Seyler 375, 419-23 (1994). Other examples of secondary metabolites are polyols, amino acids, polyketides, non-ribosomal polypeptides, ergosterol, carotenoids, terpenoids, sterols, vitamin E, and the like.

### 8.7.6 INCREASE ABILITY TO SEPARATE IN ETHANOL

Another desired property is to increase the flocculence of yeast to facilitate separation in preparation of ethanol. Yeast can be stochastic &/or non-stochastic mutagenized by any of the procedures noted above with selection for stochastic &/or non-stochastic mutagenized yeast forming the largest clumps.

### 8.7.6.1 EXEMPLARY PROCEDURE FOR YEAST PROTOPLASTING

Protoplast preparation in yeast is reviewed by Morgan, in *Protoplasts* (Birkhauser Verlag, Basel, 1983). Fresh cells ( $\sim 10^8$ ) are washed with buffer, for example 0.1 M potassium phosphate, then resuspended in this same buffer containing a reducing agent, such as 50 mM DTT, incubated for 1 h at 30°C with gentle agitation, and then washed again with buffer to remove the reducing agent. These cells are then resuspended in buffer containing a cell wall degrading enzyme, such as Novozyme 234 (1 mg/mL), and any of a variety of osmotic stabilizers, such as sucrose, sorbitol, NaCl, KCl,  $\text{MgSO}_4$ ,  $\text{MgCl}_2$ , or  $\text{NH}_4\text{Cl}$  at any of a variety of concentrations. These suspensions are then incubated at 30°C with gentle shaking ( $\sim 60$  rpm) until protoplasts are released. To generate protoplasts that are more likely to produce productive fusants several strategies are possible.

Protoplast formation can be increased if the cell cycle of the protoplasts have been synchronized to be halted at G1. In the case of *S. cerevisiae* this can be accomplished by the addition of mating factors, either  $\alpha$  or  $a$  (Curran & Carter, *J Gen. Microbiol.* 129, 1589-1591 (1983)). These peptides act as adenylate cyclase inhibitors which by decreasing the cellular level of cAMP arrest the cell cycle at G1. In addition, sex factors have been shown to induce the weakening of the cell wall in preparation for the sexual fusion of  $\alpha$  and  $a$  cells (Crandall & Brock, *Bacteriol. Rev.* 32, 139-163 (1968); Osumi et al., *Arch. Microbiol.* 97, 27-38 (1974)). Thus in the preparation of protoplasts, cells can be treated with mating factors or other known inhibitors of adenylate cyclase, such as leflunomide or the killer toxin from *K. lactis*, to arrest them at G1 (Sugisaki et al., *Nature* 304, 464-466 (1983)). Then after fusing of the protoplasts (step 2), cAMP can be added to the regeneration medium to induce S-phase and DNA synthesis. Alternatively, yeast strains having a temperature sensitive mutation in the *CDC4* gene can be used, such that cells could be synchronized and arrested at G1. After fusion cells are returned to the permissive temperature so that DNA synthesis and growth resumes.

Once suitable protoplasts have been prepared, it is necessary to induce fusion by physical or chemical means. An equal number of protoplasts of each cell type is mixed in

phosphate buffer (0.2 M, pH 5.8,  $2 \times 10^8$  cells/mL) containing an osmotic stabilizer, for example 0.8 M NaCl, and PEG 6000 (33% w/v) and then incubated at 30°C for 5 min while fusion occurs. Polyols, or other compounds that bind water, can be employed. The fusants are then washed and resuspended in the osmotically stabilized buffer lacking PEG, and transferred to osmotically stabilized regeneration medium on/in which the cells can be selected or screened for a desired property.

#### **8.7.7 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS USING ARTIFICIAL CHROMOSOMES**

Yeast artificial chromosomes (Yacs) are yeast vectors into which very large DNA fragments (e.g., 50-2000 kb) can be cloned (see, e.g., Monaco & Larin, Trends. Biotech. 12(7), 280-286 (1994); Ramsay, Mol Biotechnol 1(2), 181-201 1994; Huxley, Genet. Eng. 16, 65-91 (1994); Jakobovits, Curr. Biol. 4(8), 761-3 (1994); Lamb & Gearhart, Curr. Opin. Genet. Dev. 5(3), 342-8 (1995); Montoliu et al., Reprod Fertil. Dev. 6, 577-84 (1994)). These vectors have telomeres (Tel), a centromere (Cen), an autonomously replicating sequence (ARS), and can have genes for positive (e.g., TRPI) and negative (e.g., URA3) selection. YACs are maintained, replicated, and segregate as other yeast chromosomes through both meiosis and mitosis thereby providing a means to expose cloned DNA to true meiotic recombination.

YACs provide a vehicle for the stochastic &/or non-stochastic mutagenesis of libraries of large DNA fragments in vivo. The substrates for stochastic &/or non-stochastic mutagenesis are typically large fragments from 20 kb to 2 Mb. The fragments can be random fragments or can be fragments known to encode a desirable property. For example, a fragment might include an operon of genes involved in production of antibiotics. Libraries can also include whole genomes or chromosomes. Viral genomes and some bacterial genomes can be cloned intact into a single YAC. In some libraries, fragments are obtained from a single organism. Other libraries include fragment variants, as where some libraries are obtained from different individuals or species. Fragment variants can also be generated by induced mutation. Typically, genes within fragments are

expressed from naturally associated regulatory sequences within yeast. However, alternatively, individual genes can be linked to yeast regulatory elements to form an expression cassette, and a concatemer of such cassettes, each containing a different gene, can be inserted into a YAC.

In some instances, fragments are incorporated into the yeast genome, and stochastic &/or non-stochastic mutagenesis is used to evolve improved yeast strains. In other instances, fragments remain as components of YACs throughout the stochastic &/or non-stochastic mutagenesis process, and after acquisition of a desired property, the YACs are transferred to a desired recipient cell.

#### **8.7.8 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF GENES FOR BIOREMEDIATION**

Modern industry generates many pollutants for which the environment can no longer be considered an infinite sink. Naturally occurring microorganisms are able to metabolize thousands of organic compounds, including many not found in nature (e.g xenobiotics). Bioremediation, the deliberate use of microorganisms for the biodegradation of man-made wastes, is an emerging technology that offers cost and practicality advantages over traditional methods of disposal. The success of bioremediation depends on the availability of organisms that are able to detoxify or mineralize pollutants.

Microorganisms capable of degrading specific pollutants can be generated by genetic engineering and recursive sequence recombination. Although bioremediation is an aspect of pollution control, a more useful approach in the long term is one of prevention before industrial waste is pumped into the environment. Exposure of industrial waste streams to recursive sequence recombination-generated microorganisms capable of degrading the pollutants they contain would result in detoxification of mineralization of these pollutants before the waste stream enters the environment. Issues of releasing recombinant organisms can be avoided by containing them within bioreactors fitted to the industrial effluent pipes. This approach would also allow the microbial mixture used to be

adjusted to best degrade the particular wastes being produced. Finally, this method would avoid the problems of adapting to the outside world and dealing with competition that face many laboratory microorganisms.

In the wild, microorganisms have evolved new catabolic activities enabling them to exploit pollutants as nutrient sources for which there is no competition. However, pollutants that are present at low concentrations in the environment may not provide a sufficient advantage to stimulate the evolution of catabolic enzymes. For a review of such naturally occurring evolution of biodegradative pathways and the manipulation of some of microorganisms by classical techniques, see Ramos et al., *Bio/Technology* 12:1349-1355 (1994).

Generation of new catabolic enzymes or pathways for bioremediation has thus relied upon deliberate transfer of specific genes between organisms (Wackett et al., *supra*), forced matings between bacteria with specific catabolic capabilities (Brenner et al. *Biodegradation* 5:359-377 (1994)), or prolonged selection in a chemostat. Some researchers have attempted to facilitate evolution via naturally occurring genetic mechanisms in their chemostat selections by including microorganisms with a variety of catabolic pathways (Kellogg et. al. *Science* 214:1133-1135 (1981); Chakrabarty *American Society of Micro. Biol. News* 62:130-137 (1996)). For a review of efforts in this area, see Cameron et al. *Applied Biochem. Biotech* 38:105-140 (1993).

Current efforts in improving organisms for bioremediation take a labor-intensive approach in which many parameters are optimized independently, including transcription efficiency from native and heterologous promoters, regulatory circuits and translational efficiency as well as improvement of protein stability and activity (Timmis et al. *Ann. Rev. Microbiol.* 48:525- 527 (1994)).

A recursive sequence recombination approach overcomes a number of limitations in the bioremediation capabilities of naturally occurring microorganisms. Both enzyme

activity and specificity can be altered, simultaneously or sequentially, by the methods of the invention. For example, catabolic enzymes can be evolved to increase the rate at which they act on a substrate. Although knowledge of a rate-limiting step in a metabolic pathway is not required to practice the invention, rate-limiting proteins in pathways can be evolved to have increased expression and/or activity, the requirement for inducing substances can be eliminated, and enzymes can be evolved that catalyze novel reactions.

Some examples of chemical targets for bioremediation include but are not limited to benzene, xylene, and toluene, camphor, naphthalene, halogenated hydrocarbons, polychlorinated biphenyls (PCBs), trichlorethylene, pesticides such as pentachlorophenyls (PCPs), and herbicides such as atrazine.

#### 8.7.8.1 AROMATIC HYDROCARBONS

Preferably, when an enzyme is "evolved" to have a new catalytic function, that function is expressed, either constitutively or in response to the new substrate. Recursive sequence recombination subjects both structural and regulatory elements (including the structure of regulatory proteins) of a protein to recombinogenic mutagenesis simultaneously. Selection of mutants that are efficiently able to use the new substrate as a nutrient source will be sufficient to ensure that both the enzyme and its regulation are optimized, without detailed analysis of either protein structure or operon regulation.

Examples of aromatic hydrocarbons include but are not limited to benzene, xylene, toluene, biphenyl, and polycyclic aromatic hydrocarbons such as pyrene and naphthalene. These compounds are metabolized via catechol intermediates. Degradation of catechol by *Pseudomonas putida* requires induction of the catabolic operon by *cis*, *cis*-muconate which acts on the CatR regulatory protein. The binding site for the CatR protein is G-N<sub>11</sub>-A, while the optimal sequence for the LysR class of activators (of which CatR is a member) is T-N<sub>11</sub>-A. Mutation of the G to a T in the CatR binding site enhances the expression of catechol metabolizing genes (Chakrabarty, American Society of Microbiology News 62:130-137 (1996)). This demonstrates that the control of existing catabolic pathways is not optimized for the metabolism of specific xenobiotics. It is also an example of a type of

mutant that would be expected from recursive sequence recombination of the operon followed by selection of bacteria that are better able to degrade the target compound.

As an example of starting materials, dioxygenases are required for many pathways in which aromatic compounds are catabolized. Even small differences in dioxygenase sequence can lead to significant differences in substrate specificity (Furukawa et al. *J. Bact.* 175:5224-5232 (1993); Erickson et al. *App. Environ. Micro.* 59:3858-3862 (1993)). A hybrid enzyme made using sequences derived from two "parental" enzymes may possess catalytic activities that are intermediate between the parents (Erickson, *ibid.*), or may actually be better than either parent for a specific reaction (Furukawa et al. *J. Bact.* 176:2121-2123 (1994)). In one of these cases site directed mutagenesis was used to generate a single polypeptide with hybrid sequence (Erickson, *ibid.*); in the other, a four subunit enzyme was produced by expressing two subunits from each of two different dioxygenases (Furukawa, *ibid.*). Thus, sequences from one or more genes encoding dioxygenases can be used in the recursive sequence recombination techniques of the instant invention, to generate enzymes with new specificities. In addition, other features of the catabolic pathway can also be evolved using these techniques, simultaneously or sequentially, to optimize the metabolic pathway for an activity of interest.

#### 8.7.8.2 HALOGENATED HYDROCARBONS

Large quantities of halogenated hydrocarbons are produced annually for uses as solvents and biocides. These include, in the United States alone, over 5 million tons of both 1,2-dichloroethane and vinyl chloride used in PVC production in the U.S. alone. The compounds are largely not biodegradable by processes in single organisms, although in principle haloaromatic catabolic pathways can be constructed by combining genes from different microorganisms. Enzymes can be manipulated to change their substrate specificities. Recursive sequence recombination offers the possibility of tailoring enzyme specificity to new substrates without needing detailed structural analysis of the enzymes.



As an example of possible starting materials for the methods of the instant invention, Wackett et al. (Nature 368:627-629 (1994)) recently demonstrated that through classical techniques a recombinant *Pseudomonas* strain in which seven genes encoding two multi-component oxygenases are combined, generated a single host that can metabolize polyhalogenated compounds by sequential reductive and oxidative techniques to yield non-toxic products. These and/or related materials can be subjected to the techniques discussed above so as to evolve and optimize a biodegradative pathway in a single organism.

Trichloroethylene is a significant groundwater contaminant. It is degraded by microorganisms in a cometabolic way (i.e., no energy or nutrients are derived). The enzyme must be induced by a different compound (e.g., *Pseudomonas cepacia* uses toluene-4-monooxygenase, which requires induction by toluene, to destroy trichloroethylene). Furthermore, the degradation pathway involves formation of highly reactive epoxides that can inactivate the enzyme (Timmis et al. Ann. Rev. Microbiol. 48:525-557 (1994)). The recursive sequence recombination techniques of the invention could be used to mutate the enzyme and its regulatory region such that it is produced constitutively, and is less susceptible to epoxide inactivation. In some embodiments of the invention, selection of hosts constitutively producing the enzyme and less susceptible to the epoxides can be accomplished by demanding growth in the presence of increasing concentrations of trichloroethylene in the absence of inducing substances.

#### **8.7.8.3 POLYCHLORINATED BIPHENYLS AND POLYCYCLIC AROMATIC HYDROCARBONS**

Polychlorinated Biphenyls (PCBs) and Polycyclic Aromatic Hydrocarbons (PAHs) are families of structurally related compounds that are major pollutants at many Superfund sites. Bacteria transformed with plasmids encoding enzymes with broader substrate specificity have been used commercially. In nature, no known pathways have been generated in a single host that degrade the larger PAHs or more heavily chlorinated PCBs. Indeed, often the collaboration of anaerobic and aerobic bacteria are required for

complete metabolism.

Thus, likely sources for starting material for recursive sequence recombination include identified genes encoding PAH-degrading catabolic pathways on large (20-100KB) plasmids (Sanseverino et al. *Applied Environ. Micro.* 59:1931-1937 (1993); Simon et al. *Gene* 127:31-37 (1993); Zylstra et al. *Annals of the NY Acad. Sci.* 721:386-398 (1994)); while biphenyl and PCB-metabolizing enzymes are encoded by chromosomal gene clusters, and in a number of cases have been cloned onto plasmids (Hayase et al. *J. Bacteriol.* 172:1160-1164 (1990); Furukawa et al. *Gene* 98:21-28 (1992); Hofer et al. *Gene* 144:9-16 (1994)). The materials can be subjected to the techniques discussed above so as to evolve a biodegradative pathway in a single organism.

Substrate specificity in the PCB pathway largely results from enzymes involved in initial dioxygenation reactions, and can be significantly altered by mutations in those enzymes (Erickson et al. *Applied Environ. Micro.* 59:3858-3866 (1993); Furukawa et al. *J. Bact.* 175:5224-5232 (1993). Mineralization of PAHs and PCBs requires that the downstream pathway is able to metabolize the products of the initial reaction (Brenner et al. *Biodegradation* 5:359-377 (1994)). In this case, recursive sequence recombination of the entire pathway with selection for bacteria able to use the PCE or PAH as the sole carbon source will allow production of novel PCB and PAH degrading bacteria.

#### 8.7.8.4 HERBICIDES

A general method for evolving genes for the catabolism of insoluble herbicides is exemplified as follows for atrazine. Atrazine [2-chloro-4-(ethylamino)-6-(isopropylamino)-1,3,5-triazine] is a moderately persistent herbicide which is frequently detected in ground and surface water at concentrations exceeding the 3 ppb health advisory level set by the EPA. Atrazine can be slowly metabolized by a *Pseudomonas* species (Mandelbaum et al. *Appl. Environ. Micro.* 61:1451-1457 (1995)). The enzymes catalyzing the first two steps in atrazine metabolism by *Pseudomonas* are encoded by genes *AtzA* and *AtzB* (de Souza et al. *Appl. Environ. Micro.* 61:3373-3378 (1995)). These genes have been cloned in a 6.8 kb fragment into pUC18 (*AtzAB*-pUC). *E. coli* carrying

this plasmid converts atrazine to much more soluble metabolites. It is thus possible to screen for enzyme activity by growing bacteria on plates containing atrazine. The herbicide forms an opaque precipitate in the plates, but cells containing AtzAB-pU18 secrete atrazine degrading enzymes, leading to a clear halo around those cells or colonies. Typically, the size of the halo and the rate of its formation can be used to assess the level of activity so that picking colonies with the largest halos allows selection of the more active or highly produced atrazine degrading enzymes.

Thus, the plasmids carrying these genes can be subjected to the recursive sequence recombination formats described above to optimize the catabolism of atrazine in *E. coli* or another host of choice, including *Pseudomonas*. After each round of recombination, screening of host colonies expressing the evolved genes can be done on agar plates containing atrazine to observe halo formation. This is a generally applicable method for screening enzymes that metabolize insoluble compounds to those that are soluble (e.g., polycyclic aromatic hydrocarbons). Additionally, catabolism of atrazine can provide a source of nitrogen for the cell; if no other nitrogen is available, cell growth will be limited by the rate at which the cells can catabolize nitrogen. Cells able to utilize atrazine as a nitrogen source can thus be selected from a background of non-utilizers or poor-utilizers.

#### 8.7.8.5 HEAVY METAL DETOXIFICATION

Bacteria are used commercially to detoxify arsenate waste generated by the mining of arsenopyrite gold ores. As well as mining effluent, industrial waste water is often contaminated with heavy metals (e.g., those used in the manufacture of electronic components and plastics). Thus, simply to be able to perform other bioremedial functions, microorganisms must be resistant to the levels of heavy metals present, including mercury, arsenate, chromate, cadmium, silver, etc.

A strong selective pressure is the ability to metabolize a toxic compound to one less toxic. Heavy metals are toxic largely by virtue of their ability to denature proteins (Ford et al. Bioextraction and Biodeterioration of Metals, p. 1-23). Detoxification of heavy

metal contamination can be effected in a number of ways including changing the solubility or bioavailability of the metal, changing its redox state (e.g. toxic mercuric chloride is detoxified by reduction to the much more volatile elemental mercury) and even by bioaccumulation of the metal by immobilized bacteria or plants. The accumulation of metals to a sufficiently high concentration allows metal to be recycled; smelting burns off the organic part of the organism, leaving behind reusable accumulated metal. Resistances to a number of heavy metals (arsenate, cadmium, cobalt, chromium, copper, mercury, nickel, lead, silver, and zinc) are plasmid encoded in a number of species including *Staphylococcus* and *Pseudomonas* (Silver et al. *Environ. Health Perspect.* 102:107-113 (1994); Ji et al. *J. Ind. Micro.* 14:61-75 (1995)). These genes also confer heavy metal resistance on other species as well (e.g., *E. coli*). The recursive sequence recombination techniques of the instant invention (RSR) can be used to increase microbial heavy metal tolerances, as well as to increase the extent to which cells will accumulate heavy metals. For example, the ability of *E. coli* to detoxify arsenate can be improved at least 100-fold by RSR.

Cyanide is very efficiently used to extract gold from rock containing as little as 0.2 oz per ton. This cyanide can be microbially neutralized and used as a nitrogen source by fungi or bacteria such as *Pseudomonas fluorescens*. A problem with microbial cyanide degradation is the presence of toxic heavy metals in the leachate. RSR can be used to increase the resistance of bioremedial microorganisms to toxic heavy metals, so that they will be able to survive the levels present in many industrial and Superfund sites. This will allow them to biodegrade organic pollutants including but not limited to aromatic hydrocarbons, halogenated hydrocarbons, and biocides.

#### 8.7.8.6 MICROBIAL MINING

"Bioleaching" is the process by which microbes convert insoluble metal deposits (usually metal sulfides or oxides) into soluble metal sulfates. Bioleaching is commercially important in the mining of arsenopyrite, but has additional potential in the detoxification and recovery of metals and acids from waste dumps. Naturally occurring bacteria capable

of bioleaching are reviewed by Rawlings and Silver (Bio/Technology 13:773-778 (1995)).

These bacteria are typically divided into groups by their preferred temperatures for growth. The more important mesophiles are *Thiobacillus* and *Leptospirillum* species. Moderate thermophiles include *Sulfobacillus* species. Extreme thermophiles include *Sulfolobus* species. Many of these organisms are difficult to grow in commercial industrial settings, making their catabolic abilities attractive candidates for transfer to and optimization in other organisms such as *Pseudomonas*, *Rhodococcus*, *T. ferrooxidans* or *E. coli*. Genetic systems are available for at least one strain of *T. ferrooxidans*, allowing the manipulation of its genetic material on plasmids.

The recursive sequence recombination methods described above can be used to optimize the catalytic abilities in native hosts or heterologous hosts for evolved bioleaching genes or pathways, such as the ability to convert metals from insoluble to soluble salts. In addition, leach rates of particular ores can be improved as a result of, for example, increased resistance to toxic compounds in the ore concentrate, increased specificity for certain substrates, ability to use different substrates as nutrient sources, and so on.

#### 8.7.8.7 OIL DESULFURIZATION

The presence of sulfur in fossil fuels has been correlated with corrosion of pipelines, pumping, and refining equipment, and with the premature breakdown of combustion engines. Sulfur also poisons many catalysts used in the refining of fossil fuels. The atmospheric emission of sulfur combustion products is known as acid rain.

Microbial desulfurization is an appealing bioremediation application. Several bacteria have been reported that are capable of catabolizing dibenzothiophene (DBT), which is the representative compound of the class of sulfur compounds found in fossil fuels. U.S. Patent No. 5,356,801 discloses the cloning of a DNA molecule from *Rhodococcus rhodochrous* capable of biocatalyzing the desulfurization of oil. Denome et

al. (Gene 175:6890-6901 (1995)) disclose the cloning of a 9.8 kb DNA fragment from *Pseudomonas* encoding the upper naphthalene catabolizing pathway which also degrades dibenzothiophene. Other genes have been identified that perform similar functions (disclosed in U.S. 5,356,801).

The activity of these enzymes is currently too low to be commercially viable, but the pathway could be increased in efficiency using the recursive sequence recombination techniques of the invention. The desired property of the genes of interest is their ability to desulfurize dibenzothiophene or its alkyl or aryl substituted analogues. In some embodiments of the invention, selection is preferably accomplished by coupling this pathway to one providing a nutrient to the bacteria. Thus, for example, desulfurization of dibenzothiophene results in formation of hydroxybiphenyl.

This is a substrate for the biphenyl-catabolizing pathway which provides carbon and energy. Selection would thus be done by "stochastic &/or non-stochastic mutagenesis" the dibenzothiophene genes and transforming them into a host containing the biphenyl-catabolizing pathway. Increased dibenzothiophene desulfurization will result in increased nutrient availability and increased growth rate. Once the genes have been evolved they are easily separated from the biphenyl degrading genes. The latter are undesirable in the final product since the object is to desulfurize without decreasing the energy content of the oil. Alkyl or aryl substituted dibenzothiophenes can be detected by changes in fluorescence (Krawiec, S., *Devel. Indus. Microbiology* 31:103-114 (1990)) or by detection of phenol groups formed as a result of desulfurization (Dacre, J.C. *Anal. Chem.* 43:589-591 (1971)).

#### 8.7.8.8 ORGANO-NITRO COMPOUNDS

Organo-nitro compounds are used as explosives, dyes, drugs, polymers and antimicrobial agents. Biodegradation of these compounds occurs usually by way of reduction of the nitrate group, catalyzed by nitroreductases, a family of broadly-specific enzymes. Partial reduction of organo-nitro compounds often results in the formation of a compound more toxic than the original (Hassan et al. 1979 *Arch Bioch Biop.* 196:385-395). Recursive sequence recombination of nitroreductases can produce enzymes that are

more specific, and able to more completely reduce (and thus detoxify) their target compounds (examples of which include but are not limited to nitrotoluenes and nitrobenzenes). Nitro-reductases can be isolated from bacteria isolated from explosive-contaminated soils, such as *Morganella morganii* and *Enterobacter cloacae* (Bryant et. al., 1991. J. Biol Chem. 266:4126-4130). A preferred selection method is to look for increased resistance to the organo-nitro compound of interest, since that will indicate that the enzyme is also able to reduce any toxic partial reduction products of the original compound.

#### 8.7.8.9 ALTERNATIVE SUBSTRATES FOR CHEMICAL SYNTHESIS

Metabolic engineering can be used to alter microorganisms that produce industrially useful chemicals, so that they will grow using alternate and more abundant sources of nutrients, including human- produced industrial wastes. This typically involves providing both a transport system to get the alternative substrate into the engineered cells and catabolic enzymes from the natural host organisms to the engineered cells.

In some instances, enzymes can be secreted into the medium by engineered cells to degrade the alternate substrate into a form that can more readily be taken up by the engineered cells; in other instances, a batch of engineered cells can be grown on one preferred substrate, then lysed to liberate hydrolytic enzymes for the alternate substrate into the medium, while a second inoculum of the same engineered host or a second host is added to utilize the hydrolyzate.

The starting materials for recursive sequence recombination will typically be genes for utilization of a substrate or its transport. Examples of nutrient sources of interest include but are not limited to lactose, whey, galactose, mannitol, xylan, cellobiose, cellulose and sucrose, thus allowing cheaper production of compounds including but not limited to ethanol, tryptophan, rhamnolipid surfactants, xanthan gum, and polyhydroxylalkanoate. For a review of such substrates as desired target substances, see Cameron et al. (Appl. Biochem. Biotechnol. 38:105-140 (1993)). The recursive sequence recombination methods described above can be used to optimize the ability of native hosts

or heterologous hosts to utilize a substrate of interest, to evolve more efficient transport systems, to increase or alter specificity for certain substrates, and so on.

#### **8.7.8.10 MODIFICATION OF CELL PROPERTIES**

Although not strictly examples of manipulation of intermediary metabolism, recursive sequence recombination techniques can be used to improve or alter other aspects of cell properties, from growth rate to ability to secrete certain desired compounds to ability to tolerate increased temperature or other environmental stresses. Some examples of traits engineered by traditional methods include expression of heterologous proteins in bacteria, yeast, and other eukaryotic cells, antibiotic resistance, and phage resistance. Any of these traits is advantageously evolved by the recursive sequence recombination techniques of the instant invention. Examples include replacement of one nutrient uptake system (e.g. ammonia in *Methylophilus methylotrophus*) with another that is more energy efficient; expression of haemoglobin to improve growth under conditions of limiting oxygen; redirection of toxic metabolic end products to less toxic compounds; expression of genes conferring tolerance to salt, drought and toxic compounds and resistance to pathogens, antibiotics and bacteriophage, reviewed in Cameron et. al. *Appl Biochem Biotechnol*, 38:105-140 (1993).

The heterologous genes encoding these functions all have the potential for further optimization in their new hosts by existing recursive sequence recombination technology. Since these functions increase cell growth rates under the desired growth conditions, optimization of the genes by evolution simply involves recombining the DNA recursively and selecting the recombinants that grow faster with limiting oxygen, higher toxic compound concentration, or whatever is the appropriate growth condition for the parameter being improved.

Since these functions increase cell growth rates under the desired growth conditions, optimization of the genes by "evolution" can simply involve "stochastic &/or non-stochastic mutagenesis" the DNA and selecting the recombinants that grow faster with limiting oxygen, higher toxic compound concentration or whatever restrictive



condition is being overcome. Cultured mammalian cells also require essential amino acids to be present in the growth medium. This requirement could also be circumvented by expression of heterologous metabolic pathways that synthesize these amino acids (Rees et al. , Biotechnology 8:629-633 (1990). Recursive sequence recombination would provide a mechanism for optimizing the expression of these genes in mammalian cells. Once again, a preferred selection would be for cells that can grow in the absence of added amino acids.

Yet another candidate for improvement through the techniques of the invention is symbiotic nitrogen fixation. Genes involved in nodulation (nod, ndv), nitrogen reduction (nif, fix), host range determination (nod, hsp), bacteriocin production (tfx), surface polysaccharide synthesis (exo) and energy utilization (dct, hup) which have been identified (Paau, Biotech. Adv. 9:173-184 (1991)). The main function of recursive sequence recombination in this case is in improving the survival of strains that are already known to be better nitrogen fixers. These strains tend to be less good at competing with strains already present in the environment, even though they are better at nitrogen fixation. Targets for recursive sequence recombination such as nodulation and host range determination genes can be modified and selected for by their ability to grow on the new host.

Similarly any bacteriocin or energy utilization genes that will improve the competitiveness of the strain will also result in greater growth rates. Selection can simply be performed by subjecting the target genes to recursive sequence recombination and forcing the inoculant to compete with wild type nitrogen fixing bacteria. The better the nitrogen fixing bacteria grow in the new host, the more copies of their recombined genes will be present for the next round of recombination. This growth rate differentiating selection is described above in detail.

#### **8.7.8.11 BIODETECTORS / BIOSENSORS**

Bioluminescence or fluorescence genes can be used as reporters by fusing them to specific regulatory genes (Cameron et. al. Appl. Biochem Biotechnol, 38:105- 140 (1993)). A specific example is one in which the luciferase genes luxCDABE of *Vibrio*

fischeri were fused to the regulatory region of the isopropylbenzene catabolism operon from *Pseudomonas putida* RE204.

Transformation of this fusion construct into *E. coli* resulted in a strain which produced light in response to a variety of hydrophobic compound such as substituted benzenes, chlorinated solvents and naphthalene (Selifonova et. al., Appl Environ Microbiol 62:778-783 (1996)). This type of construct is useful for the detection of pollutant levels, and has the added benefit of only measuring those pollutants that are bioavailable (and therefore potentially toxic). other signal molecules such as jellyfish green fluorescent protein could also be fused to genetic regulatory regions that respond to chemicals in the environment. This should allow a variety of molecules to be detected by their ability to induce expression of a protein or proteins which result in light, fluorescence or some other easily detected signal. Recursive sequence recombination can be used in several ways to modify this type of biodetection system. It can be used to increase the amplitude of the response, for example by increasing the fluorescence of the green fluorescent protein. Recursive sequence recombination could also be used to increase induced expression levels or catalytic activities of other signal-generating systems, for example of the luciferase genes.

Recursive sequence recombination can also be used to alter the specificity of biosensors. The regulatory region, and transcriptional activators that interact with this region and with the chemicals that induce transcription can also be stochastic &/or non-stochastic mutagenized. This should generate regulatory systems in which transcription is activated by analogues of the normal inducer, so that biodetectors for different chemicals can be developed.

In this case, selection would be for constructs that are activated by the (new) specific chemical to be detected. Screening could be done simply with fluorescence (or light) activated cell sorting, since the desired improvement is in light production. In addition to detection of environmental pollutants, biosensors can be developed that will respond to any chemical for which there are receptors, or for which receptors can be

evolved by recursive sequence recombination, such as hormones, growth factors, metals and drugs. These receptors may be intracellular and direct activators of transcription, or they may be membrane bound receptors that activate transcription of the signal indirectly, for example by a phosphorylation cascade. They may also not act on transcription at all, but may produce a signal by some post-transcriptional modification of a component of the signal generating pathway. These receptors may also be generated by fusing domains responsible for binding different ligands with different signaling domains. Again, recursive sequence recombination can be used to increase the amplitude of the signal generated to optimize expression and functioning of chimeric receptors, and to alter the specificity of the chemicals detected by the receptor.

## 8.8 PROMOTING GENETIC EXCHANGE

Some methods of the invention effect recombination of cellular DNA by propagating cells under conditions inducing exchange of DNA between cells. DNA exchange can be promoted by generally applicable methods such as electroporation, biolistics, cell fusion, or in some instances, by conjugation, transduction, or agrobacterium mediated transfer and meiosis. For example, *Agrobacterium* can transform *S. cerevisiae* with T-DNA, which is incorporated into the yeast genome by both homologous recombination and a gap repair mechanism. (Piers et al., Proc. Natl. Acad. Sci. USA 93(4),1613-8 (1996)).

In some methods, initial diversity between cells (i.e., before genome exchange) is induced by chemical or radiation-induced mutagenesis of a progenitor cell type, optionally followed by screening for a desired phenotype. In other methods, diversity is natural as where cells are obtained from different individuals, strains or species.

In some stochastic &/or non-stochastic mutagenesis methods, induced exchange of DNA is used as the sole means of effecting recombination in each cycle of recombination. In other methods, induced exchange is used in combination with natural sexual recombination of an organism. In other methods, induced exchange and/or natural sexual

recombination are used in combination with the introduction of a fragment library. Such a fragment library can be a whole genome, a whole chromosome, a group of functionally or genetically linked genes, a plasmid, a cosmid, a mitochondrial genome, a viral genome (replicative and nonreplicative) or specific or random fragments of any of these. The DNA can be linked to a vector or can be in free form. Some vectors contain sequences promoting homologous or nonhomologous recombination with the host genome. Some fragments contain double stranded breaks such as caused by shearing with glass beads, sonication, or chemical or enzymatic fragmentation, to stimulate recombination. In each case, DNA can be exchanged between cells after which it can undergo recombination to form hybrid genomes. Generally, cells are recursively subject to recombination to increase the diversity of the population prior to screening. Cells bearing hybrid genomes, e.g., generated after at least one, and usually several cycles of recombination are screened for a desired phenotype, and cells having this phenotype are isolated. These cells can additionally form starting materials for additional cycles of recombination in a recursive recombination/selection scheme.

#### 8.8.1 PROTOPLAST FUSION

One means of promoting exchange of DNA between cells is by fusion of cells, such as by protoplast fusion. A protoplast results from the removal from a cell of its cell wall, leaving a membrane-bound cell that depends on an isotonic or hypertonic medium for maintaining its integrity. If the cell wall is partially removed, the resulting cell is strictly referred to as a spheroplast and if it is completely removed, as a protoplast. However, here the term protoplast includes spheroplasts unless otherwise indicated.

Protoplast fusion is described by Shaffner et al., Proc. Natl. Acad. Sci. USA 77, 2163 (1980) and other exemplary procedures are described by Yoakum et al., US 4,608,339, Takahashi et al., US 4,677,066 and Sambrooke et al., at Ch. 16. Protoplast fusion has been reported between strains, species, and genera (e.g., yeast and chicken erythrocyte). Protoplasts can be prepared for both bacterial and eukaryotic cells, including mammalian cells and plant cells, by several means including chemical treatment to strip

cell walls. For example, cell walls can be stripped by digestion with a cell wall degrading enzyme such as lysozyme in a 10-20% sucrose, 50 mM EDTA buffer. Conversion of cells to spherical protoplasts can be monitored by phase-contrast microscopy. Protoplasts can also be prepared by propagation of cells in media supplemented with an inhibitor of cell wall synthesis, or use of mutant strains lacking capacity for cell wall formation.

Preferably, eukaryotic cells are synchronized in G1 phase by arrest with inhibitors such as - factor, *K. lactis* killer toxin, leflonamide and adenylate cyclase inhibitors. Optionally, some but not all, protoplasts to be fused can be killed and/or have their DNA fragmented by treatment with ultraviolet irradiation, hydroxylamine or cupferon (Reeves et al., *FFMS Microbiol. Lett.* 99, 193 - 198 (1992)). In this situation, killed protoplasts are referred to as donors, and viable protoplasts as acceptors.

Using dead donors cells can be advantageous in subsequently recognizing fused cells with hybrid genomes, as described below. Further, breaking up DNA in donor cells is advantageous for stimulating recombination with acceptor DNA. Optionally, acceptor and/or fused cells can also be briefly, but nonlethally, exposed to UV irradiation further to stimulate recombination.

Once formed, protoplasts can be stabilized in a variety of osmolytes and compounds such as sodium chloride, potassium chloride, sodium phosphate, potassium phosphate, sucrose, sorbitol in the presence of DTT. The combination of buffer, pH, reducing agent, and osmotic stabilizer can be optimized for different cell types. Protoplasts can be induced to fuse by treatment with a chemical such as PEG, calcium chloride or calcium propionate or electrofusion (Tsoneva, *Acta Microbiologica Bulgaria* 24, 53-59 (1989)). A method of cell fusion employing electric fields has also been described. See Chang US, 4,970,154. Conditions can be optimized for different strains.

The fused cells are heterokaryons containing genomes from two or more component protoplasts. Fused cells can be enriched from unfused parental cells by sucrose gradient sedimentation or cell sorting. The two nuclei in the heterokaryons can fuse (karyogamy) and homologous recombination can occur between the genomes. The

chromosomes can also segregate asymmetrically resulting in regenerated protoplasts that have lost or gained whole chromosomes. The frequency of recombination can be increased by treatment with ultraviolet irradiation or by use of strains overexpressing *recA* or other recombination genes, or the yeast *rad* genes, and cognate variants thereof in other species, or by the inhibition of gene products of *MutS*, *MutL*, or *MutD*. Overexpression can be either the result of introduction of exogenous recombination genes or the result of selecting strains, which as a result of natural variation or induced mutation, overexpress endogenous recombination genes. The fused protoplasts are propagated under conditions allowing regeneration of cell walls, recombination and segregation of recombinant genomes into progeny cells from the heterokaryon and expression of recombinant genes. This process can be reiteratively repeated to increase the diversity of any set of protoplasts or cells. After, or occasionally before or during, recovery of fused cells, the cells are screened or selected for evolution toward a desired property.

Thereafter a subsequent round of recombination can be performed by preparing protoplasts from the cells surviving selection/screening in a previous round. The protoplasts are fused, recombination occurs in fused protoplasts, and cells are regenerated from the fused protoplasts. This process can again be reiteratively repeated to increase the diversity of the starting population. Protoplasts, regenerated or regenerating cells are subject to further selection or screening.

Subsequent rounds of recombination can be performed on a split pool basis as described above. That is, a first subpopulation of cells surviving selection/screening from a previous round are used for protoplast formation. A second subpopulation of cells surviving selection/screening from a previous round are used as a source for DNA library preparation.

The DNA library from the second subpopulation of cells is then transformed into the protoplasts from the first subpopulation. The library undergoes recombination with the genomes of the protoplasts to form recombinant genomes. This process can be repeated several times in the absence of a selection event to increase the diversity of the cell

population. Cells are regenerated from protoplasts, and selection/screening is applied to regenerating or regenerated cells. In a further variation, a fresh library of nucleic acid fragments is introduced into protoplasts surviving selection/screening from a previous round.

Protoplast formation of donor and recipient strains, heterokaryon formation, karyogamy, recombination, and segregation of recombinant genomes into separate cells. Optionally, the recombinant genomes, if having a sexual cycle, can undergo further recombination with each other as a result of meiosis and mating. Recursive cycles of protoplast fusion, or recursive mating/meiosis is often used to increase the diversity of a cell population. After achieving a sufficiently diverse population via one of these forms of recombination, cells are screened or selected for a desired property. Cells surviving selection/screening can then be used as the starting materials in a further cycle of protoplasting or other recombination methods as noted herein.

#### 8.8.2 PARASEXUAL REPRODUCTION

Parasexual reproduction provides a further means for stochastic &/or non-stochastic mutagenesis genetic material between cells. This process allows recombination of parental DNA without involvement of mating types or gametes. Parasexual fusion occurs by hyphal fusion giving rise to a common cytoplasm containing different nuclei. The two nuclei can divide independently in the resulting heterokaryon but occasionally fuse. Fusion is followed by haploidization, which can involve loss of chromosomes and mitotic crossing over between homologous chromosomes. Protoplast fusion is a form of parasexual reproduction.

### **8.8.3 SELECTION FOR HYBRID STRAINS**

#### **8.8.3.1 IDENTIFYING CELLS FORMED BY THE FUSION OF COMPONENTS OF PARENTAL CELLS FROM TWO OR MORE DISTINCT SUBPOPULATIONS**

The invention provides selection strategies to identify cells formed by fusion of components from parental cells from two or more distinct subpopulations. Selection for hybrid cells is usually performed before selecting or screening for cells that have evolved (as a result of genetic exchange) to acquisition of a desired property. A basic premise of most such selection schemes is that two initial subpopulations have two distinct markers. Cells with hybrid genomes can thus be identified by selection for both markers.

#### **8.8.3.2 METHOD WHERE ONE SUBPOPULATION HAS A MARKER**

In one such scheme, at least one subpopulation of cells bears a selective marker attached to its cell membrane. Examples of suitable membrane markers include biotin, fluorescein and rhodamine. The markers can be linked to amide or thiol groups or through more specific derivatization chemistries, such as iodo-acetates, iodoacetamides, maleimides.

For example, a marker can be attached as follows. Cells or protoplasts are washed with a buffer (e.g., PBS), which does not interfere with the chemical coupling of a chemically active ligand which reacts with amino groups of lysines or N-terminal amino groups of membrane proteins. The ligand is either amine reactive itself (e.g., isothiocyanates, succinimidyl esters, sulfonyl chlorides) or is activated by a heterobifunctional linker (e.g. EMCS, SLAB, SPDP, SMB) to become amine reactive. The ligand is a molecule which is easily bound by protein derivatized magnetic beads or other capturing solid supports. For example, the ligand can be succinimidyl activated biotin (Molecular Probes Inc.: B-1606, B-2603, S-1515, S-1582). This linker is reacted with amino groups of proteins residing in and on the surface of a cell. The cells are then washed to remove excess labeling agent before contacting with cells from the second subpopulation bearing a second selective marker.



The second subpopulation of cells can also bear a membrane marker, albeit a different membrane marker from the first subpopulation. Alternatively, the second subpopulation can bear a genetic marker. The genetic marker can confer a selective property such as drug resistance or a screenable property, such as expression of green fluorescent protein.

After fusion of first and second subpopulations of cells and recovery, cells are screened or selected for the presence of markers on both parental subpopulations. For example, fusants are enriched for one population by adsorption to specific beads and these are then sorted by FACS for those expressing a marker. Cells surviving both screens for both markers are those having undergone protoplast fusion, and are therefore more likely to have recombined genomes. Usually, the markers are screened or selected separately. Membrane-bound markers, such as biotin, can be screened by affinity enrichment for the cell membrane marker (e.g., by panning fused cells on an affinity matrix). For example, for a biotin membrane label, cells can be affinity purified using streptavidin-coated magnetic beads (Dynal). These beads are washed several times to remove the non-fused host cells.

Alternatively, cells can be panned against an antibody to the membrane marker. In a further variation, if the membrane marker is fluorescent, cells bearing the marker can be identified by FACS. Screens for genetic markers depend on the nature of the markers, and include capacity to grow on drug-treated media or FACS selection for green fluorescent protein. If first and second cell populations have fluorescent markers of different wavelengths, both markers can be screened simultaneously by FACS sorting.

In a further selection scheme for hybrid cells, first and second populations of cells to be fused express different subunits of a heteromultimeric enzyme. Usually, the heteromultimeric enzyme has two different subunits, but heteromultimeric enzymes having three, four or more different subunits can be used. If an enzyme has more than two different subunits, each subunit can be expressed in a different subpopulation of cells (e.g.,

three subunits in three subpopulations), or more than one subunit can be expressed in the same subpopulation of cells (e.g., one subunit in one subpopulation, two subunits in a second subpopulation). In the case where more than two subunits are used, selection for the poolwise recombination of more than two protoplasts can be achieved.

Hybrid cells representing a combination of genomes of first, second or more subpopulation component cells can then be recognized by an assay for intact enzyme. Such an assay can be a binding assay, but is more typically a functional assay (e.g., capacity to metabolize a substrate of the enzyme). Enzymatic activity can be detected for example by processing of a substrate to a product with a fluorescent or otherwise easily detectable absorbance or emission spectrum. The individual subunits of a heteromultimeric enzyme used in such an assay preferably have no enzymic activity in dissociated form, or at least have significantly less activity in dissociated form than associated form. Preferably, the cells used for fusion lack an endogenous form of the heteromultimeric enzyme, or at least have significantly less endogenous activity than results from heteromultimeric enzyme formed by fusion of cells.

Penicillin acylase enzymes, cephalosporin acylase and penicillin acyltransferase are examples of suitable heteromultimeric enzymes. These enzymes are encoded by a single gene, which is translated as a proenzyme and cleaved by posttranslational autocatalytic proteolysis to remove a spacer endopeptide and generate two subunits, which associate to form the active heterodimeric enzyme. Neither subunit is active in the absence of the other subunit. However, activity can be reconstituted if these separated gene portions are expressed in the same cell by co-transformation. Other enzymes that can be used have subunits that are encoded by distinct genes (e.g., *faoA* and *faoB* genes encode 3-oxoacyl-CoA thiolase of *Pseudomonas fragi* (Biochem. J 328, 815-820 (1997))).

An exemplary enzyme is penicillin G acylase from *Escherichia coli*, which has two subunits encoded by a single gene. Fragments of the gene encoding the two subunits operably linked to appropriate expression regulation sequences are transfected into first and second subpopulations of cells, which lack endogenous penicillin acylase activity. A

cell formed by fusion of component cells from the first and second subpopulations expresses the two subunits, which assemble to form functional enzyme, e.g., penicillin acylase. Fused cells can then be selected on agar plates containing penicillin G, which is degraded by penicillin acylase.

In another variation, fused cells are identified by complementation of auxotrophic mutants. Parental subpopulations of cells can be selected for known auxotrophic mutations. Alternatively, auxotrophic mutations in a starting population of cells can be generated spontaneously by exposure to a mutagenic agent. Cells with auxotrophic mutations are selected by replica plating on minimal and complete media. Lesions resulting in auxotrophy are expected to be scattered throughout the genome, in genes for amino acid, nucleotide, and vitamin biosynthetic pathways. After fusion of parental cells, cells resulting from fusion can be identified by their capacity to grow on minimal media. These cells can then be screened or selected for evolution toward a desired property. Further steps of mutagenesis generating fresh auxotrophic mutations can be incorporated in subsequent cycles of recombination and screening/selection.

In variations of the above method, de novo generation of auxotrophic mutations in each round of stochastic &/or non-stochastic mutagenesis can be avoided by reusing the same auxotrophs. For example, auxotrophs can be generated by transposon mutagenesis using a transposon bearing selective marker. Auxotrophs are identified by a screen such as replica plating. Auxotrophs are pooled, and a generalized transducing phage lysate is prepared by growth of phage on a population of auxotrophic cells. A separate population of auxotrophic cells is subjected to genetic exchange, and complementation is used to select cells that have undergone genetic exchange and recombination. These cells are then screened or selected for acquisition of a desired property. Cells surviving screening or selection then have auxotrophic markers regenerated by introduction of the transducing transposon library. The newly generated auxotrophic cells can then be subject to further genetic exchange and screening/selection.

In a further variation, auxotrophic mutations are generated by homologous recombination with a targeting vector comprising a selective marker flanked by regions of homology with a biosynthetic region of the genome of cells to be evolved. Recombination between the vector and the genome inserts the positive selection marker into the genome causing an auxotrophic mutation. The vector is in linear form before introduction of cells.

Optionally, the frequency of introduction of the vector can be increased by capping its ends with self-complementarity oligonucleotides annealed in a hair pin formation. Genetic exchange and screening/selection proceed as described above. In each round, targeting vectors are reintroduced regenerating the same population of auxotrophic markers.

In another variation, fused cells are identified by screening for a genomic marker present on one subpopulation of parental cells and an episomal marker present on a second subpopulation of cells. For example, a first subpopulation of yeast containing mitochondria can be used to complement a second subpopulation of yeast having a petite phenotype (i.e., lacking mitochondria).

In a further variation, genetic exchange is performed between two subpopulations of cells, one of which is dead. Cells are preferably killed by brief exposure to DNA fragmenting agents such as hydroxylamine, cupferon, or irradiation. Viable cells are then screened for a marker present on the dead parental subpopulation.

#### **8.8.4 LIPOSOME MEDIATED TRANSFERS**

##### **8.8.4.1 NUCLEIC ACID FRAGMENT LIBRARIES ARE INTRODUCED INTO PROTOPLASTS**

###### **8.8.4.1.1 THE NUCLEIC ACIDS ARE ENCAPSULATED IN LIPOSOMES TO HELP UPTAKE BY PROTOPLASTS**

In the methods noted above, in which nucleic acid fragment libraries are introduced into protoplasts, the nucleic acids are sometimes encapsulated in liposomes to

facilitate uptake by protoplasts. Liposome-mediated uptake of DNA by protoplasts is described in Redford et al., *Mol. Gen. Genet.* 184, 567-569 (1981). Liposomes can efficiently deliver large volumes of DNA to protoplasts (see Deshayes et al., *FMBO J.* 4, 2731-2737 (1985)).

See also, Philippot and Schuber (eds) (1995) *Liposomes as Tools in Basic Research and Industry* CRC press, Boca Raton, e.g., Chapter 9, Remy et al. "Gene Transfer with Cationic Amphiphiles." Further, the DNA can be delivered as linear fragments, which are often more recombinogenic than whole genomes. In some methods, fragments are mutated prior to encapsulation in liposomes. In some methods, fragments are combined with RecA and homologs, or nucleases (e.g., restriction endonucleases) before encapsulation in liposomes to promote recombination. Alternatively, protoplasts can be treated with lethal doses of nicking reagents and then fused. Cells which survive are those which are repaired by recombination with other genomic fragments, thereby providing a selection mechanism to select for recombinant (and therefore desirably diverse) protoplasts.

## 8.9 SHUFFLING USING FILAMENTOUS FUNGI

Filamentous fungi are particularly suited to performing the stochastic &/or non-stochastic mutagenesis methods described above. Filamentous fungi are divided into four main classifications based on their structures for sexual reproduction: Phycomycetes, Ascomycetes, Basidiomycetes and the Fungi Imperfecti. Phycomycetes (e.g., *Rhizopus*, *Mucor*) form sexual spores in sporangium.

The spores can be uni or multinucleate and often lack septated hyphae (coenocytic).

Ascomycetes (e.g., *Aspergillus*, *Neurospora*, *Penicillium*) produce sexual spores in an ascus as a result of meiotic division. Asci typically contain 4 meiotic products, but some contain 8 as a result of additional mitotic division. Basidiomycetes include mushrooms, and smuts and form sexual spores on the surface of a basidium. In

holobasidiomycetes, such as mushrooms, the basidium is undivided. In hemibasidiomycetes, such as rusts (Uredinales) and smut fungi (Ustilaginales), the basidium is divided. Fungi imperfecti, which include most human pathogens, have no known sexual stage.

Fungi can reproduce by asexual, sexual or parasexual means. Asexual reproduction, involves vegetative growth of mycelia, nuclear division and cell division without involvement of gametes and without nuclear fusion. Cell division can occur by sporulation, budding or fragmentation of hyphae.

#### **8.9.1 EVOLVE FUNGI FROM STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS TO BECOME USEFUL HOSTS FOR GENETIC ENGINEERING OF UNRELATED GENES**

#### **8.9.2 TO IMPROVE THE CAPACITY OF FUNGI TO MAKE SPECIFIC COMPOUNDS**

One general goal of stochastic &/or non-stochastic mutagenesis is to evolve fungi to become useful hosts for genetic engineering, in particular for the stochastic &/or non-stochastic mutagenesis of unrelated genes. *A. nidulans* and *neurospora* are generally the fungal organisms of choice to serve as a hosts for such manipulations because of their sexual cycles and well-established use in classical and molecular genetics. Another general goal is to improve the capacity of fungi to make specific compounds (e.g. antibacterials (penicillins, cephalosporins), antifungals (e.g. echinocandins, aureobasidins), and wood-degrading enzymes). There is some overlap between these general goals, and thus, some desired properties are useful for achieving both goals.

#### **8.9.3 MUTATOR STRAIN**

Another desired property is the production of a mutator strain of fungi. Such a fungus can be produced by stochastic &/or non-stochastic mutagenesis a fungal strain containing a marker gene with one or more mutations that impair or prevent expression of a functional product. Shufflants are propagated under conditions that select for expression

of the positive marker (while allowing a small amount of residual growth without expression). Shufflants growing fastest are selected to form the starting materials for the next round of stochastic &/or non-stochastic mutagenesis.

#### **8.9.4 EXPANDED HOST RANGE SO ABLE TO FORM HETEROKARYONS WITH MORE STRAINS**

Another desired property is to expand the host range of a fungus so it can form heterokaryons with fungi from other vegetative compatibility groups. Incompatibility between species results from the interactions of specific alleles at different incompatibility loci (such as the "het" loci). If two strains undergo hyphal anastomosis, a lethal cytoplasmic incompatibility reaction may occur if the strains differ at these loci. Strains must carry identical loci to be entirely compatible. Several of these loci have been identified in various species, and the incompatibility effect is somewhat additive (hence, "partial incompatibility" can occur). Some tolerant and het-negative mutants have been described for these organisms (e.g. Dales & Croft, *J Gen. Microbiol.* 136, 1717-1724 (1990)). Further, a tolerance gene (*tol*) has been reported, which suppresses mating-type heterokaryon incompatibility. Stochastic &/or non-stochastic mutagenesis is performed between protoplasts of strains from different incompatibility groups. A preferred format uses a five acceptor strain and a UV-irradiated dead acceptor strain. The UV irradiation serves to introduce mutations into DNA inactivating het genes. The two strains should bear different genetic markers. Protoplasts of the strain are fused, cells are regenerated and screened for complementation of markers. Subsequent rounds of stochastic &/or non-stochastic mutagenesis and selection can be performed in the same manner by fusing the cells surviving screening with protoplasts of a fresh population of donor cells. Similar to other procedures noted herein, the cells resulting from regeneration of the protoplasts are optionally refused by protoplasting and regenerated into cells one or more times prior to any selection step to increase the diversity of the resulting population of cells to be screened.

### 8.9.5 ABILITY TO OUTBREED WITHOUT SELF-BREEDING

Another desired property is the introduction of multiple-allelomorph heterothallism into Ascomyces and Fungi imperfecti, which do not normally exhibit this property. This mating system allows outbreeding without self-breeding. Such a mating system can be introduced by stochastic &/or non-stochastic mutagenesis Ascomycetes and Fungi imperfecti with DNA from Gasteromycetes or Hymenomycetes, which have such a system.

### 8.9.6 SPONTANEOUS FORMATION OF PROTOPLASTS

Another desired property is spontaneous formation of protoplasts to facilitate use of a fungal strain as a stochastic &/or non-stochastic mutagenesis host. Here, the fungus to be evolved is typically mutagenized. Spores of the fungus to be evolved are briefly treated with a cell-wall degrading agent for a time insufficient for complete protoplast formation, and are mixed with protoplasts from other strain(s) of fungi. Protoplasts formed by fusion of the two different subpopulations are identified by genetic or other selection/or screening as described above. These protoplasts are used to regenerate mycelia and then spores, which form the starting material for the next round of stochastic &/or non-stochastic mutagenesis. In the next round, at least some of the surviving spores are treated with cell-wall removing enzyme but for a shorter time than the previous round. After treatment, the partially stripped cells are labeled with a first label. These cells are then mixed with protoplasts, which may derive from other cells surviving selection in a previous round, or from a fresh strain of fungi. These protoplasts are physically labeled with a second label. After incubating the cells under conditions for protoplast fusion fusants with both labels are selected.

These fusants are used to generate mycelia and spores for the next round of stochastic &/or non-stochastic mutagenesis, and so forth. Eventually, progeny that spontaneously form protoplasts (i.e., without addition of cell wall degrading agent) are identified. As with other procedures noted herein, cells or protoplasts can be reiteratively fused and regenerated prior to performing any selection step to increase the diversity of



the resulting cells or protoplasts to be screened. Similarly, selected cells or protoplasts can be reiteratively fused and regenerated for one or several cycles without imposing selection on the resulting cellular or protoplast populations, thereby increasing the diversity of cells or protoplasts which are eventually screened. This process of performing multiple cycles of recombination interspersed with selection steps can be reiteratively repeated as desired.

#### **8.9.7 ACQUISITION OR IMPROVEMENT OF GENES ENCODING IN BIOSYNTHETIC PATHWAYS; TRANSPORTER PROTEINS; AND METABOLIC FLUX**

Another desired property is the acquisition and/or improvement of genes encoding enzymes in biosynthetic pathways, genes encoding transporter proteins, and genes encoding proteins involved in metabolic flux control. In this situation, genes of the pathway can be introduced into the fungus to be evolved either by genetic exchange with another strain of fungus possessing the pathway or by introduction of a fragment library from an organism possessing the pathway. Genetic material of these fungi can then be subjected to further stochastic &/or non-stochastic mutagenesis and screening/selection by the various procedures discussed in this application.

Shufflant strains of fungi are selected/screened for production of the compound produced by the metabolic pathway or precursors thereof.

#### **8.9.8 INCREASED STABILITY TO EXTREME CONDITIONS**

Another desired property is increasing the stability of fungi to extreme conditions such as heat. In this situation, genes conferring stability can be acquired by exchanging DNA with or transforming DNA from a strain that already has such properties.

Alternatively, the strain to be evolved can be subjected to random mutagenesis. Genetic material of the fungus to be evolved can be stochastic &/or non-stochastic mutagenized by any of the procedures described in this application, with shufflants being

selected by surviving exposure to extreme conditions.

### **8.9.9 GROWTH UNDER ALTERED NUTRITIONAL REQUIREMENTS**

Another desired property is capacity of a fungus to grow under altered nutritional requirements (e.g., growth on particular carbon or nitrogen sources). Altering nutritional requirements is particularly valuable, e.g., for natural isolates of fungi that produce valuable commercial products but have esoteric and therefore expensive nutritional requirement. The strain to be evolved undergoes genetic exchange and/or transformation with DNA from a strain that has the desired nutritional requirements. The fungus to be evolved can then optionally be subjected to further stochastic &/or non-stochastic mutagenesis as described in this application and with recombinant strains being selected for capacity to grow in the desired nutritional circumstances. Optionally, the nutritional circumstances can be varied in successive rounds of stochastic &/or non-stochastic mutagenesis starting at close to the natural requirements of the fungus to be evolved and in subsequent rounds approaching the desired nutritional requirements.

### **8.9.10 NATURAL COMPETANCE TO TAKE UP A PLASMID BEARING A SELECTIVE MARKER**

Another desired property is acquisition of natural competence in a fungus. The procedure for acquisition of natural competence by stochastic &/or non-stochastic mutagenesis is generally described in PCT/US97/04494. The fungus to be evolved typically undergoes genetic exchange or transformation with DNA from a bacterial strain or fungal strain that already has this property.

Cells with recombinant genomes are then selected by capacity to take up a plasmid bearing a selective marker. Further rounds of recombination and selection can be performed using any of the procedures described above.

### 8.9.11 REDUCED OR INCREASED SECRETION OF PROTEASES AND DNASES

Another desired property is reduced or increased secretion of proteases and DNase. In this situation, the fungus to be evolved can acquire DNA by exchange or transformation from another strain known to have the desired property. Alternatively, the fungus to be evolved can be subject to random mutagenesis. The fungus to be evolved is stochastic &/or non-stochastic mutagenized as above. The presence of such enzymes, or lack thereof, can be assayed by contacting the culture media from individual isolates with a fluorescent molecule tethered to a support via a peptide or DNA linkage. Cleavage of the linkage releases detectable fluorescence to the media.

### 8.9.12 ALTERED TRANSPORTERS TO USE SECONDARY COMPONENTS

Another desired property is producing fungi with altered transporters (e. g., MDR). Such altered transporters are useful, for example, in fungi that have been evolved to produce new secondary metabolites, to allow entry of precursors required for synthesis of the new secondary metabolites into a cell, or to allow efflux of the secondary metabolite from the cell. Transporters can be evolved by introduction of a library of transporter variants into fungal cells and allowing the cells to recombine by sexual or parasexual recombination. To evolve a transporter with capacity to transport a precursor into the cells, cells are propagated in the presence of precursor, and cells are then screened for production of metabolite. To evolve a transporter with capacity to export a metabolite, cells are propagated under conditions supporting production of the metabolite, and screened for export of metabolite to culture medium.

A general method of fungal stochastic &/or non-stochastic mutagenesis is shown herein. Spores from a frozen stock, a lyophilized stock, or fresh from an agar plate are used to inoculate suitable liquid medium (1). Spores are germinated resulting in hyphal growth (2). Mycelia are harvested, and washed by filtration and/or centrifugation. Optionally the sample is pretreated with DTT to enhance protoplast formation (3). Protoplasting is performed in an osmotically stabilizing medium (e.g., 1 M NaCl/20mM MgSO<sub>4</sub>, pH 5.8) by the addition of cell wall- degrading enzyme (e.g., Novozyme 234) (4).

Cell wall degrading enzyme is removed by repeated washing with osmotically stabilizing solution (5). Protoplasts can be separated from mycelia, debris and spores by filtration through miracloth, and density centrifugation (6). Protoplasts are harvested by centrifugation and resuspended to the appropriate concentration. This step may lead to some protoplast fusion (7). Fusion can be stimulated by addition of PEG (e.g., PEG 3350), and/or repeated centrifugation and resuspension with or without PEG. Electrofusion can also be performed (8). Fused protoplasts can optionally be enriched from unfused protoplasts by sucrose gradient sedimentation (or other methods of screening described above). Fused protoplasts can optionally be treated with ultraviolet irradiation to stimulate recombination (9). Protoplasts are cultured on osmotically stabilized agar plates to regenerate cell walls and form mycelia (10). The mycelia are used to generate spores (11), which are used as the starting material in the next round of stochastic &/or non-stochastic mutagenesis (12). Selection for a desired property can be performed either on regenerated mycelia or spores derived therefrom.

In an alternative method, protoplasts are formed by inhibition of one or more enzymes required for cell wall synthesis. The inhibitor should be fungistatic rather than fungicidal under the conditions of use. Examples of inhibitors include antifungal compounds described by (e.g., Georgopapadakou & Walsh, *Antimicrob. Ag. Chemother.* 40, 279-291 (1996); Lyman & Walsh, *Drugs* 44, 9-35 (1992)). Other examples include chitin synthase inhibitors (polyoxin or nikkomycin compounds) and/or glucan synthase inhibitors (e.g. echinocandins, papulocandins, pneumocandins). Inhibitors should be applied in osmotically stabilized medium. Cells stripped of their cell walls can be fused or otherwise employed as donors or hosts in genetic transformation/strain development programs.

In a further variation, protoplasts are prepared using strains of fungi, which are genetically deficient or compromised in their ability to synthesize intact cell walls. Such mutants are generally referred to as fragile, osmotic-remedial, or cell wall-less, and are obtainable from strain depositories. Examples of such strains include *Neurospora crassa* os mutants (Selitrennikoff, *Antimicrob. Agents. Chemother.* 23, 757-765 (1983)). Some such

mutations are temperature-sensitive. Temperature-sensitive strains can be propagated at the permissive temperature for purposes of selection and amplification and at a nonpermissive temperature for purposes of protoplast formation and fusion. A temperature sensitive strain *Neurospora crassa* os strain has been described which propagates as protoplasts when growth in osmotically stabilizing medium containing sorbose and polyoxin at nonpermissive temperature but generates whole cells on transfer to medium containing sorbitol at a permissive temperature. See US 4,873,196.

Other suitable strains can be produced by targeted mutagenesis of genes involved in chitin synthesis, glucan synthesis and other cell wall-related processes. Examples of such genes include CHT1, CHT2 and CAL1 (or CSD2) of *Saccharomyces cerevisiae* and *Candida* spp. (Georgopapadakou & Walsh 1996); ETG1/FKSI/CND1/CWH53/PB RI and homologs in *S. cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, *Aspergillus fumigatus*, ChvAINDvA *Agrobacterium* and *Rhizobium*. Other examples are M4, orlB, orlC, MD, tsE, and bimG of *Aspergillus nidulans* (Borgia, J Bacteriol. 174, 3 77-3 89 (1992)).

Strains of *A. nidulans* containing Or1A1 or tse1 mutations lyse at restrictive temperatures. Lysis of these strains may be prevented by osmotic stabilization, and the mutations may be complemented by the addition of N-acetylglucosimine (GlcNac). BimG11 mutations are ts for a type 1 protein phosphatase (germlines of strains carrying this mutation lack chitin, and conidia swell and lyse). Other suitable genes are chsA, chsB, chsC, chsD and chsE of *Aspergillus fumigatus*; chs1 and chs2 of *Neurospora crassa*; *Phycomyces blakesleeanus* MM and chs 1, 2 and 3 of *S. cerevisiae*. Chs 1 is a non-essential repair enzyme; chs2 is involved in septum formation and chs3 is involved in cell wall maturation and bud ring formation.

Other useful strains include *S. cerevisiae* CLY (cell lysis) mutants such as ts strains (Paravicini et al., Mol. Cell Biol 12, 4896-4905 (1992)), and the CLY 15 strain which harbors a PKC 1 gene deletion. Other useful strains include strain VY 1160 containing a ts mutation in *srb* (encoding actin) (Schade et al. Acta Histochem. Suppl. 41, 193-200

(1991)), and a strain with an *ses* mutation which results in increased sensitivity to cell-wall digesting enzymes isolated from snail gut (Metha & Gregory, *Appl. Environ. Microbiol.* 41, 992-999 (1981)). Useful strains of *C. albicans* include those with mutations in *chs1*, *chs2*, or *chs3* (encoding chitin synthetases), such as osmotic remedial conditional lethal mutants described by Payton & de Tiani, *Curr. Genet.* 17, 293-296 (1990); *C. utilis* mutants with increased sensitivity to cell-wall digesting enzymes isolated from snail gut (Metha & Gregory, 1981, *supra*); and *X. crassa* mutants *os-1*, *os-2*, *os-3*, *os-4*, *os-5*, and *os-6*. See, Selitrennikoff, *Antimicrob. Agents Chemother.* 23, 757-765 (1983). Such mutants grow and divide without a cell wall at 37 °C, but at 22 °C produce a cell wall.

Targeted mutagenesis can be achieved by transforming cells with a positive-negative selection vector containing homologous regions flanking a segment to be targeted, a positive selection marker between the homologous regions and a negative selection marker outside the homologous regions (see Capecchi, US 5,627,059). In a variation, the negative selection marker can be an antisense transcript of the positive selection marker (see US 5,527,674).

Other suitable cells can be selected by random mutagenesis or stochastic &/or non-stochastic mutagenesis procedures in combination with selection. For example, a first subpopulation of cells are mutagenized, allowed to recover from mutagenesis, subjected to incomplete degradation of cell walls and then contacted with protoplasts of a second subpopulation of cells. Hybrid cells bearing markers from both subpopulations are identified (as described above) and used as the starting materials in a subsequent round of stochastic &/or non-stochastic mutagenesis. This selection scheme selects both for cells with capacity for spontaneous protoplast formation and for cells with enhanced recombinationogenicity.

In a further variation, cells having capacity for spontaneous protoplast formation can be crossed with cells having enhanced recombinationogenicity evolved using other methods of the invention. The hybrid cells are particularly suitable hosts for whole genome stochastic &/or non-stochastic mutagenesis.

Cells with mutations in enzymes involved in cell wall synthesis or maintenance can undergo fusion simply as a result of propagating the cells in osmotic-protected culture due to spontaneous protoplast formation. If the mutation is conditional, cells are shifted to a nonpermissive condition. Protoplast formation and fusion can be accelerated by addition of promoting agents, such as PEG or an electric field (See Philipova & Venkov, Yeast 6, 205-212 (1990); Tsoneva et al., FFMS Microbiol Lett. 51, 61-65 (1989)).

## 8.10 PROCESS OF SEXUAL REPRODUCTION

Sexual reproduction provides a mechanism for stochastic &/or non-stochastic mutagenesis genetic material between cells. A sexual reproductive cycle is characterized by an alteration of a haploid phase and a diploid phase. Diploidy occurs when two haploid gamete nuclei fuse (karyogamy). The gamete nuclei can come from the same parental strains (self-fertile), such as in the homothallic fungi. In heterothallic fungi, the parental strains come from strains of different mating type.

A diploid cell converts to haploidy via meiosis, which essentially consists of two divisions of the nucleus accompanied by one division of the chromosomes. The products of one meiosis are a tetrad (4 haploid nuclei). In some cases, a mitotic division occurs after meiosis, giving rise to eight product cells. The arrangement of the resultant cells (usually enclosed in spores) resembles that of the parental strains. The length of the haploid and diploid stages differs in various fungi: for example, the Basidiomycetes and many of the Ascomycetes have a mostly haploid life cycle (that is, meiosis occurs immediately after karyogamy), whereas others (e.g., *Saccharomyces cerevisiae*) are diploid for most of their life cycle (karyogamy occurs soon after meiosis). Sexual reproduction can occur between cells in the same strain (selfing) or between cells from different strains (outcrossing).

Sexual dimorphism (dioecism) is the separate production of male and female organs on different mycelia. This is a rare phenomenon among the fungi, although a few examples are known. Heterothallism (one locus-two alleles) allows for outcrossing between crosscompatible strains which are self-incompatible. The simplest form is the

two allele-one locus system of mating types/factors, illustrated by the following organisms: A and a in *Neurospora*; a and  $\alpha$  in *Saccharomyces*, plus and minus in *Schizosaccharomyces* and *Zygomycetes*;  $\alpha_1$  and  $\alpha_2$  in *Ustilago*.

Multiple-allelomorph heterothallism is exhibited by some of the higher Basidiomycetes (e.g. *Gasteromycetes* and *Hymenomycetes*), which are heterothallic and have several mating types determined by multiple alleles. Heterothallism. In these organisms is either bipolar with one mating type factor, or tetrapolar with two unlinked factors, A and B. Stable, fertile heterokaryon formation depends on the presence of different A factors and, in the case of tetrapolar organisms, of different B factors as well. This system is effective in the promotion of outbreeding and the prevention of self-breeding. The number of different mating factors may be very large (i.e. thousands) (Kothe, FEMS Microbiol Rev. 18, 65-87 (1996)), and non-parental mating factors may arise by recombination.

#### **8.10.1 INTRODUCING SEXUAL CYCLES**

#### **8.10.2 MEIOSIS**

##### **8.10.2.1 HETEROKARYON-A CELL OR HYPHA CONTAINING TWO OR MORE NUCLEI OF DIFFERENT GENETIC CONSTITUTIONS**

One desired property is the introduction of meiotic apparatus into fungi presently lacking a sexual cycle (see Sharon et al., Mol. Gen. Genet. 251, 60-68 (1996)). A scheme for introducing a sexual cycle into the fungi *P. chrysogenum* (a fungus imperfecti) is shown herein. Subpopulations of protoplasts are formed from *A. nidulans* (which has a sexual cycle) and *P. chrysogenum*, which does not. The two strains preferably bear different markers. The *A. nidulans* protoplasts are killed by treatment with UV or hydroxylamine. The two subpopulations are fused to form heterokaryons. In some heterokaryons, nuclei fuse, and some recombination occurs. Fused cells are cultured under conditions to generate new cell walls and then to allow sexual recombination to occur.



Cells with recombinant genomes are then selected (e.g., by selecting for complementation of auxotrophic markers present on the respective parent strains). Cells with hybrid genomes are more likely to have acquired the genes necessary for a sexual cycle. Protoplasts of cells can then be crossed with killed protoplasts of a further population of cells known to have a sexual cycle (the same or different as the previous round) in the same manner, followed by selection for cells with hybrid genomes.

#### 8.10.2.2 VEGETATIVE COMPATIBILITY BETWEEN CLASSES OF FUNGI

Within the above four classes, fungi are also classified by vegetative compatibility group. Fungi within a vegetative compatibility group can form heterokaryons with each other. Thus, for exchange of genetic material between different strains of fungi, the fungi are usually prepared from the same vegetative compatibility group. However, some genetic exchange can occur between fungi from different incompatibility groups as a result of parasexual reproduction (see Timberlake et al., US 5,605,820). Further, as discussed elsewhere, the natural vegetative compatibility group of fungi can be expanded as a result of stochastic &/or non-stochastic mutagenesis.

Several isolates of *Aspergillus nidulans*, *A. flavus*, *A. fumigatus*, *Penicillium chrysogenum*, *P. notatum*, *Cephalosporium chrysogenum*, *Neurospora crassa*, *Aureobasidium pullulans* have been karyotyped. Genome sizes generally range between 20 and 50 Mb among the *Aspergilli*. Differences in karyotypes often exist between similar strains and are also caused by transformation with exogenous DNA. Filamentous fungal genes contain introns, usually ~50-100 bp in size, with similar consensus 5' and 3' splice sequences. Promotion and termination signals are often cross-recognizable, enabling the expression of a gene/pathway from one fungus (e.g. *A. nidulans*) in another (e.g. *P. chrysogenum*). The major components of the fungal cell wall are chitin (or chitosan), beta-glucan, and mannoproteins. Chitin and beta-glucan form the scaffolding, mannoproteins are interstitial components which dictate the wall's porosity, antigenicity and adhesion. Chitin synthetase catalyzes the polymerization of beta-(1,4)-linked N-acetylglucosamine (GlcNAc) residues, forming linear strands running antiparallel; beta-

(1,3)-glucan synthetase catalyze the homopolymerization of glucose.

## **8.11 EVOLUTION**

### **8.11.1 ARTIFICIALLY EVOLVING CELLS TO ACQUIRE A NEW OR IMPROVED PROPERTY BY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

The invention provides a number of strategies for evolving metabolic and bioprocessing pathways through the technique of recursive sequence recombination. One strategy entails evolving genes that confer the ability to use a particular substrate of interest as a nutrient source in one species to confer either more efficient use of that substrate in that species, or comparable or more efficient use of that substrate in a second species. Another strategy entails evolving genes that confer the ability to detoxify a compound of interest in one or more species of organisms. Another strategy entails evolving new metabolic pathways by evolving an enzyme or metabolic pathway for biosynthesis or degradation of a compound A related to a compound B for the ability to biosynthesize or degrade compound B, either in the host of origin or a new host. A further strategy entails evolving a gene or metabolic pathway for more efficient or optimized expression of a particular metabolite or gene product. A further strategy entails evolving a host/vector system for expression of a desired heterologous product. These strategies may involve using all the genes in a multi-step pathway, one or several genes, genes from different organisms, or one or more fragments of a gene.

The strategies generally entail evolution of gene(s) or segment(s) thereof to allow retention of function in a heterologous cell or improvement of function in a homologous or heterologous cell. Evolution is effected generally by a process termed recursive sequence recombination. Recursive sequence recombination can be achieved in many different formats and permutations of formats, as described in further detail below. These formats share some common principles. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity, i.e., the creation of a family of nucleic acid molecules showing substantial sequence identity to each other but differing in

the presence of mutations. Each recombination cycle is followed by at least one cycle of screening or selection for molecules having a desired characteristic. The molecule(s) selected in one round form the starting materials for generating diversity in the next round. In any given cycle, recombination can occur in vivo or in vitro. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying prior methods of mutagenesis (e.g., error-prone PCR or cassette mutagenesis, passage through bacterial mutator strains, treatment with chemical mutagens) to either the substrates for or products of recombination.

## **8.11.2 BASIC APPROACH**

### **8.11.2.1 SUCCESSIVE CYCLES OF RECOMBINATION AND SCREENING/SELECTION**

The invention provides methods for artificially evolving cells to acquire a new or improved property by recursive sequence recombination. Briefly, recursive sequence recombination entails successive cycles of recombination to generate molecular diversity and screening/selection to take advantage of that molecular diversity. That is, a family of nucleic acid molecules is created showing substantial sequence and/or structural identity but differing as to the presence of mutations. These sequences are then recombined in any of the described formats so as to optimize the diversity of mutant combinations represented in the resulting recombined library. Typically, any resulting recombinant nucleic acids or genomes are recursively recombined for one or more cycles of recombination to increase the diversity of resulting products. After this recursive recombination procedure, the final resulting products are screened and/or selected for a desired trait or property.

Alternatively, each recombination cycle can followed by at least one cycle of screening or selection for molecules having a desired characteristic. In this embodiment, the molecule(s) selected in one round form the starting materials for generating diversity in the next round.

The cells to be evolved can be bacteria, archaeobacteria, or eukaryotic cells and can constitute a homogeneous cell line or mixed culture. Suitable cells for evolution include the bacterial and eukaryotic, cell lines commonly used in genetic engineering, protein expression, or the industrial production or conversion of proteins, enzymes, primary metabolites, secondary metabolites, fine, specialty or commodity chemicals. Suitable mammalian cells include those from, e.g., mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIH3T3), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean, sugarcane, tobacco, and arabidopsis; fish, algae, fungi (penicillium, aspergillus, podosporea, neurospora, saccharomyces), insect (e.g., baculo lepidoptera), yeast (picchia and saccharomyces, Schizosaccharomyces pombe). Also of interest are many bacterial cell types, both gram-negative and gram-positive, such as *Bacillus subtilis*, *B. licheniformis*, *B. cereus*, *Escherichia coli*, *Streptomyces*, *Pseudomonas*, *Salmonella*, *Actinomycetes*, *Lactobacillus*, *Acetobacter*, *Deinococcus*, and *Erwinia*. The complete genome sequences of *E. coli* and *Bacillus subtilis* are described by Blattner et al., *Science* 277, 1454- 1462 (1997); Kunst et al., *Nature* 390, 249-256 (1997)

#### 8.11.2.1.1 GOAL IS TO ACHIEVE VARIATION

Evolution commences by generating a population of variant cells. Typically, the cells in the population are of the same type but represent variants of a progenitor cell. In some instances, the variation is natural as when different cells are obtained from different individuals within a species, from different species or from different genera. In other instances, variation is induced by mutagenesis of a progenitor cell. Mutagenesis can be effected by subjecting the cell to mutagenic agents, or if the cell is a mutator cell (e.g., has mutations in genes involved in DNA replication, recombination and/or repair which favor introduction of mutations) simply by propagating the mutator cells. Mutator cells can be generated from successive selections for simple phenotypic changes (e.g., acquisition of rifampicin-resistance, then nalidixic acid resistance then lac<sup>-</sup> to lac<sup>+</sup> (see Mao et al., J

Bacteriol 179, 417-422 (1997)), or mutator cells can be generated by exposure to specific inhibitors of cellular factors that result in the mutator phenotype. These could be inhibitors of mutS, mutL, mutD, recD, mutY, mutM, dam, uvrD and the like.

More generally, mutations are induced in cell populations using any available mutation technique. Common mechanisms for inducing mutations include, but are not limited to, the use of strains comprising mutations such as those involved in mismatch repair. e.g. mutations in mutS, mutT, mutL and mutH; exposure to UV light; Chemical mutagenesis, e.g. use of inhibitors of MMR, DNA damage inducible genes, or SOS inducers; overproduction/ underproduction/ mutation of any component of the homologous recombination complex/pathway, e.g. RecA, ssb, etc. overproduction/ underproduction/ mutation of genes involved in DNA synthesis/homeostasis; overproduction/ underproduction/ mutation of recombination-stimulating genes from bacteria, phage (e.g. Lambda Red function), or other organisms; addition of chi sites into/flanking the donor DNA fragments; coating the DNA fragments with RecA/ssb and the like.

In other instances, variation is the result of transferring a library of DNA fragments into the cells (e.g., by conjugation, protoplast fusion, liposome fusion, transformation, transduction or natural competence). At least one, and usually many of the fragments in the library, show some, but not complete, sequence or structural identity with a cognate or allelic gene within the cells sufficient to allow homologous recombination to occur.

For example, in one embodiment, homologous integration of a plasmid carrying a stochastic &/or non-stochastic mutagenized gene or metabolic pathway leads to insertion of the plasmid-borne sequences adjacent to the genomic copy. Optionally, a counter-selectable marker strategy is used to select for recombinants in which recombination occurred between the homologous sequences, leading to elimination of the counter-selectable marker. A variety of selectable and counter selectable markers are amply illustrated in the art. For a list of useful markers, see, Berg and Berg (1996), Transposable element tools for microbial genetics, Escherichia coli and Salmonella Neidhardt.

Washington, D.C., ASM Press. 2: 2588-2612; La Rossa, *ibid.*, 2527-2587. This strategy can be recursively repeated to maximize sequence diversity of targeted genes prior to screening/ selection for a desired trait or property.

The library of fragments can derive from one or more sources. One source of fragments is a genomic library of fragments from a different species, cell type, organism or individual from the cells being transfected. In this situation, many of the fragments in the library have a cognate or allelic gene in the cells being transformed but differ from that gene due to the presence of naturally occurring species variation, polymorphisms, mutations, and the presence of multiple copies of some homologous genes in the genome. Alternatively, the library can be derived from DNA from the same cell type as is being transformed after that DNA has been subject to induced mutation, by conventional methods, such as radiation, error-prone PCR, growth in a mutator organism, transposon mutagenesis, or cassette mutagenesis.

Alternatively, the library can derive from a genomic library of fragments generated from the pooled genomic DNA of a population of cells having the desired characteristics. Alternatively, the library can derive from a genomic library of fragments generated from the pooled genomic DNA of a population of cells having desired characteristics.

In any of these situations, the genomic library can be a complete genomic library or subgenome & library deriving, for example, from a selected chromosome, or part of a chromosome or an episomal element within a cell. As well as, or instead of these sources of DNA fragments, the library can contain fragments representing natural or selected variants of selected genes of known function (i.e., focused libraries).

The number of fragments in a library can vary from a single fragment to about  $10^{10}$  with libraries having from  $10^3$  to  $10^8$  fragments being common. The fragments should be sufficiently long that they can undergo homologous recombination and sufficiently short that they can be introduced into a cell, and if necessary, manipulated before introduction. Fragment sizes can range from about 10 b to about 20mb. Fragments can be double- or single-stranded. The fragments can be introduced into cells as whole genomes or as

components of viruses, plasmids, YACS, HACs or BACs or can be introduced as they are, in which case all or most of the fragments lack an origin of replication. Use of viral fragments with single-stranded genomes offer the advantage of delivering fragments in single stranded form, which promotes recombination. The fragments can also be joined to a selective marker before introduction. Inclusion of fragments in a vector having an origin of replication affords a longer period of time after introduction into the cell in which fragments can undergo recombination with a cognate gene before being degraded or selected against and lost from the cell, thereby increasing the proportion of cells with recombinant genomes. Optionally, the vector is a suicide vector capable of a longer existence than an isolated DNA fragment but not capable of permanent retention in the cell line. Such a vector can transiently express a marker for a sufficient time to screen for or select a cell bearing the vector (e.g., because cells transduced by the vector are the target cell type to be screened in subsequent selection assays), but is then degraded or otherwise rendered incapable of expressing the marker. The use of such vectors can be advantageous in performing optional subsequent rounds of recombination to be discussed below. For example, some suicide vectors express a long-lived toxin which is neutralized by a short-lived molecule expressed from the same vector. Expression of the toxin alone will not allow vector to be established. Jense & Gerdes, *Mol. Microbiol.* 17, 205-210 (1995); Bernard et al., *Gene* 162, 159-160. Alternatively, a vector can be rendered suicidal by incorporation of a defective origin of replication (e.g. a temperature-sensitive origin of replication) or by omission of an origin of replication. Vectors can also be rendered suicidal by inclusion of negative selection markers, such as *ura3* in yeast or *sacB* in many bacteria.

These genes become toxic only in the presence of specific compounds. Such vectors can be selected to have a wide range of stabilities. A list of conditional replication defects for vectors which can be used, e.g., to render the vector replication defective is found, e.g., in Berg and Berg (1996), "Transposable element tools for microbial genetics" *Escherichia coli* and *Salmonella* Neidhardt. Washington, D.C., ASM Press. 2: 2588-2612. Similarly, a list of counter selectable markers, generally applicable to vector selection is also found in Berg and Berg, *id.* See also, LaRossa (1996), "Mutant selections linking

physiology, inhibitors, and genotypes" *Escherichia coli* and *Salmonella* F. C. Neidhardt. Washington, D.C., ASM Press. 2: 2527-2587.

After introduction into cells, the fragments can recombine with DNA present in the genome, or episomes of the cells by homologous, nonhomologous or site-specific recombination. For present purposes, homologous recombination makes the most significant contribution to evolution of the cells because this form of recombination amplifies the existing diversity between the DNA of the cells being transfected and the DNA fragments. For example, if a DNA fragment being transfected differs from a cognate or allelic gene at two positions, there are four possible recombination products, and each of these recombination products can be formed in different cells in the transformed population. Thus, homologous recombination of the fragment doubles the initial diversity in this gene. When many fragments recombine with corresponding cognate or allelic genes, the diversity of recombination products with respect to starting products increases exponentially with the number of mutations. Recombination results in modified cells having modified genomes and/or episomes. Recursive recombination prior to selection further increases diversity of resulting modified cells.

The variant cells, whether the result of natural variation, mutagenesis, or recombination are screened or selected to identify a subset of cells that have evolved toward acquisition of a new or improved property. The nature of the screen, of course, depends on the property and several examples will be discussed below. Typically, recombination is repeated before initial screening. Optionally, however, the screening can also be repeated before performing subsequent cycles of recombination. Stringency can be increased in repeated cycles of screening. The subpopulation of cells surviving screening are optionally subjected to a further round of recombination. In some instances, the further round of recombination is effected by propagating the cells under conditions allowing exchange of DNA between cells. For example, protoplasts can be formed from the cells, allowed to fuse, and regenerated. Cells with recombinant genomes are propagated from the fused protoplasts. Alternatively, exchange of DNA can be promoted by propagation of



cells or protoplasts in an electric field. For cells having a conjugative transfer apparatus, exchange of DNA can be promoted simply by propagating the cells.

#### **8.11.2.1.2 USING TWO SEPARATE POOLS: AMPLIFY FIRST POOL AND ADD TO SECOND POOL**

In other methods, the further round of recombination is performed by a split and pool approach. That is, the surviving cells are divided into two pools. DNA is isolated from one pool, and if necessary amplified, and then transformed into the other pool.

Accordingly, DNA fragments from the first pool constitute a further library of fragments and recombine with cognate fragments in the second pool resulting in further diversity. As shown, a pool of mutant bacteria with improvements in a desired phenotype is obtained and split. Genes are obtained from one half, e.g., by PCR, by cloning of random genomic fragments, by infection with a transducing phage and harvesting transducing particles, or by the introduction of an origin of transfer (OriT) randomly into the relevant chromosome to create a donor population of cells capable of transferring random fragments by conjugation to an acceptor population. These genes are then stochastic &/or non-stochastic mutagenized (in vitro by known methods or in vivo as taught herein), or simply cloned into an allele replacement vector (e.g., one carrying selectable and counter-selectable markers). The gene pool is then transformed into the other half of the original mutant pool and recombinants are selected and screened for further improvements in phenotype. These best variants are used as the starting point for the next cycle. Alternatively, recursive recombination by any of the methods noted can be performed prior to screening, thereby increasing the diversity of the population of cells to be screened.

#### **8.11.2.1.3 SURVIVING CELLS ARE TRANSFECTED INTO FRESH DNA**

In other methods, some or all of the cells surviving screening are transfected with a fresh library of DNA fragments, which can be the same or different from the library used in the first round of recombination. In this situation, the genes in the fresh library undergo

recombination with cognate genes in the surviving cells. If genes are introduced as components of a vector, compatibility of this vector with any vector used in a previous round of transfection should be considered. If the vector used in a previous round was a suicide vector, there is no problem of incompatibility. If, however, the vector used in a previous round was not a suicide vector, a vector having a different incompatibility origin should be used in the subsequent round. In all of these formats, further recombination generates additional diversity in the DNA component of the cells resulting in further modified cells.

The further modified cells are subjected to another round of screening/selection according to the same principles as the first round. Screening/selection identifies a subpopulation of further modified cells that have further evolved toward acquisition of the property. This subpopulation of cells can be subjected to further rounds of recombination and screening according to the same principles, optionally with the stringency of screening being increased at each round. Eventually, cells are identified that have acquired the desired property.

### **8.11.3 VARIATIONS**

#### **8.11.3.1 COATING WITH RecA TO ENRICH DIVERSITY OF HOMOLOGOUS RECOMBINATION**

The frequency of homologous recombination between library fragments and cognate endogenous genes can be increased by coating the fragments with a recombinogenic protein before introduction into cells. See Pati et al., *Molecular Biology of Cancer* 1, 1 (1996); Sena & Zarling, *Nature Genetics* 3, 365 (1996); Revet et al., *J Mol. Biol.* 232, 779- 791 (1993); Kowalczkowski & Zarling in *Gene Targeting* (CRC 1995), Ch. 7. The recombinogenic protein promotes homologous pairing and/or strand exchange. The best characterized recA protein is from *E. coli* and is available from Pharmacia (Piscataway, NJ).

In addition to the wild-type protein, a number of mutant *recA*-like proteins have been identified (e.g., *recA803*). Further, many organisms have *recA*-like recombinases with strand-transfer activities (e.g., Ogawa et al., Cold Spring Harbor Symposium on Quantitative Biology 18, 567-576 (1993); Johnson & Symington, Mol. Cell. Biol. 15, 4843-4850 (1995); Fugisawa et al., Nucl. Acids Res. 13, 7473 (1985); Hsieh et al., Cell 44, 885 (1986); Hsieh et al., J Biol. Chem. 264, 5089 (1989); Fishel et al., Proc. Natl. Acad. Sci. USA 85, 3683 (1988); Cassuto et al., Mol. Gen. Genet. 208, 10 (1987); Ganea et al., Mol. Cell Biol. 7, 3124 (1987); Moore et al., J Biol. Chem. 19, 11108 (1990); Keene et al., Nucl. Acids Res. 12, 3057 (1984); Kimeic, Cold Spring Harbor Symp. 48, 675 (1984); Kimeic, Cell 44, 545 (1986); Kolodner et al., Proc. Natl. Acad. Sci. USA 84, 5560 (1987); Sugino et al., Proc. Natl. Acad. Sci. USA 85, 3683 (1988). Halbrook et al., J. Biol. Chem. 264, 21403 (1989); Eisen et al., Proc. Natl. Acad. Sci. USA 85, 7481 (1988); McCarthy et al., Proc. Natl. Acad. Sci. USA 85, 5854 (1988). Lowenhaupt et al., J Biol. Chem. 264, 20568 (1989). Examples of such recombinase proteins include *recA*, *recA803*, *uvsX*, (Roca, A.I., Crit. Rev. Biochem. Molec. Biol. 25, 415 (1990)), *sepI* (Kolodner et al., Proc. Natl. Acad. Sci. (U.S.A.) 84, 5560 (1987); Tishkoff et al., Molec. Cell. Biol 11, 2593), *RuvC* (Dunderdale et al., Nature 354, 506 (1991)), *DS72*, *KEMI*, *XRATI* (Dykstra et al., Molec. Cell. Biol 11, 2583 (1991)), *STP /DST1* (Clark et al., Molec. Cell. Biol 11, 2576 (1991)), *HPP-I* (Moore et al., Proc. Natl. Acad. Sci. (U.S.A.) 88, 9067 (1991)), other eukaryotic recombinases (Bishop et al., Cell 69, 439 (1992); Shinohara et al., Cell 69, 457). *RecA* protein forms a nucleoprotein filament when it coats a single-stranded DNA. In this nucleoprotein filament, one monomer of *recA* protein is bound to about 3 nucleotides. This property of *recA* to coat single-stranded DNA is essentially sequence independent, although particular sequences favor initial loading of *recA* onto a polynucleotide (e.g., nucleation sequences). The nucleoprotein filament(s) can be formed on essentially any DNA to be stochastic &/or non-stochastic mutagenized and can form complexes with both single-stranded and double-stranded DNA in prokaryotic and eukaryotic cells.

Before contacting with *recA* or other recombinase, fragments are often denatured, e.g., by heat-treatment. *RecA* protein is then added at a concentration of about 1-10 gM. After incubation, the *recA*-coated single-stranded DNA is introduced into recipient cells

by conventional methods, such as chemical transformation or electroporation. In general, it can be desirable to coat the DNA with a RecA homolog isolated from the organism into which the coated DNA is being delivered. Recombination involves several cellular factors and the host RecA equivalent generally interacts better with other host factors than less closely related RecA molecules. The fragments undergo homologous recombination with cognate endogenous genes. Because of the increased frequency of recombination due to recombinase coating, the fragments need not be introduced as components of vectors. Fragments are sometimes coated with other nucleic acid binding proteins that promote recombination, protect nucleic acids from degradation, or target nucleic acids to the nucleus. Examples of such proteins includes *Agrobacterium* virE2 (Duffenberger et al., Proc. Natl. Acad. Sci. USA 86, 9154-9158 (1989)). Alternatively, the recipient strains are deficient in RecD activity. Single stranded ends can also be generated by 3'-5' exonuclease activity or restriction enzymes producing 5' overhangs.

#### **8.11.3.2 AFFINITY CHROMATOGRAPHY WITH MutS TO ENRICH FOR FRAGMENTS HAVING AT LEAST ONE MISMATCH**

The *E. coli* mismatch repair protein MutS can be used in affinity chromatography to enrich for fragments of double-stranded DNA containing at least one base of mismatch. The MutS protein recognizes the bubble formed by the individual strands about the point of the mismatch. See, e.g., Hsu & Chang, WO 9320233. The strategy of affinity enriching for partially mismatched duplexes can be incorporated into the present methods to increase the diversity between an incoming library of fragments and corresponding cognate or allelic genes in recipient cells.

MutS is used to increase diversity. The DNA substrates for enrichment are substantially similar to each other but differ at a few sites.

For example, the DNA substrates can represent complete or partial genomes (e.g., a chromosome library) from different individuals with the differences being due to polymorphisms. The substrates can also represent induced mutants of a wild type sequence.

The DNA substrates are pooled, restriction digested, and denatured to produce fragments of single-stranded DNA. The single-stranded DNA is then allowed to reanneal. Some single-stranded fragments reanneal with a perfectly matched complementary strand to generate perfectly matched duplexes. Other single-stranded fragments anneal to generate mismatched duplexes. The mismatched duplexes are enriched from perfectly matched duplexes by MutS chromatography (e.g., with MutS immobilized to beads). The mismatched duplexes recovered by chromatography are introduced into recipient cells for recombination with cognate endogenous genes as described above. MutS affinity chromatography increases the proportion of fragments differing from each other and the cognate endogenous gene. Thus, recombination between the incoming fragments and endogenous genes results in greater diversity.

A second strategy for MutS enrichment. In this strategy, the substrates for MutS enrichment represent variants of a relatively short segment, for example, a gene or cluster of genes, in which most of the different variants differ at no more than a single nucleotide. The goal of MutS enrichment is to produce substrates for recombination that contain more variations than sequences occurring in nature. This is achieved by fragmenting the substrates at random to produce overlapping fragments. The fragments are denatured and reannealed as in the first strategy. Reannealing generates some mismatched duplexes which can be separated from perfectly matched duplexes by MutS affinity chromatography. As before, MutS chromatography enriches for duplexes bearing at least a single mismatch. The mismatched duplexes are then stochastic &/or non-stochastic mutagenized into longer fragments. This is accomplished by cycles of denaturation, reannealing, and chain extension of partially annealed duplexes. After several such cycles, fragments of the same length as the original substrates are achieved, except that these fragments differ from each other at multiple sites. These fragments are then introduced into cells where they undergo recombination with cognate endogenous genes.

### 8.11.3.3 SUICIDE VECTOR ENRICHES MUTATIONS FOR CELLS THAT HAVE INTEGRATED THE VECTOR INTO THE HOST CHROMOSOME

The invention further provides methods of enriching for cells bearing modified genes relative to the starting cells. This can be achieved by introducing a DNA fragment library (e.g., a single specific segment or a whole or partial genomic library) in a suicide vector (i.e., lacking a functional replication origin in the recipient cell type) containing both positive and negative selection markers. Optionally, multiple fragment libraries from different sources (e.g., *B. subtilis*, *B. licheniformis* and *B. cereus*) can be cloned into different vectors bearing different selection markers. Suitable positive selection markers include *neo*<sup>R</sup>, *kanamycin*<sup>R</sup>, *hyg*, *hisD*, *gpt*, *ble*, *tet*<sup>R</sup>. Suitable negative selection markers include *hsv-tk*, *hprt*, *gpt*, *SacB* *ura3* and cytosine deaminase. A variety of examples of conditional replication vectors, mutations affecting vector replication, limited host range vectors, and counterselectable markers are found in Berg and Berg, *supra*, and LaRossa, *ibid.* and the references therein.

In one example, a plasmid with R6K and *fl* origins of replication, a positively selectable marker (beta-lactamase), and a counterselectable marker (*B. subtilis* *sacB*) was used. M 13 transduction of plasmids containing cloned genes were efficiently recombined into the chromosomal copy of that gene in a rep mutant *E. coli* strain.

Another strategy for applying negative selection is to include a wild type *rpsL* gene (encoding ribosomal protein S12) in a vector for use in cells having a mutant *rpsL* gene conferring streptomycin resistance. The mutant form of *rpsL* is recessive in cells having wild type *rpsL*. Thus, selection for Sm resistance selects against cells having a wild type copy of *rpsL*. See Skorupski & Taylor, *Gene* 169, 47-52 (1996). Alternatively, vectors bearing only a positive selection marker can be used with one round of selection for cells expressing the marker, and a subsequent round of screening for cells that have lost the marker (e.g., screening for drug sensitivity). The screen for cells that have lost the positive selection marker is equivalent to screening against expression of a negative selection marker. For example, *Bacillus* can be transformed with a vector bearing a CAT gene and a sequence to be integrated. See Harwood & Cutting, *Molecular Biological Methods for*

Bacillus, at pp. 31-33. Selection for chloramphenicol resistance isolates cells that have taken up vector. After a suitable period to allow recombination, selection for CAT sensitivity isolates cells which have lost the CAT gene. About 50% of such cells will have undergone recombination with the sequence to be integrated.

Suicide vectors bearing a positive selection marker and optionally, a negative selection marker and a DNA fragment can integrate into host chromosomal DNA by a single crossover at a site in chromosomal DNA homologous to the fragment. Recombination generates an integrated vector flanked by direct repeats of the homologous sequence. In some cells, subsequent recombination between the repeats results in excision of the vector and either acquisition of a desired mutation from the vector by the genome or restoration of the genome to wild type.

In the present methods, after transfer of the gene library cloned in a suitable vector, positive selection is applied for expression of the positive selection marker. Because nonintegrated copies of the suicide vector are rapidly eliminated from cells, this selection enriches for cells that have integrated the vector into the host chromosome. The cells surviving positive selection can then be propagated and subjected to negative selection, or screened for loss of the positive selection marker. Negative selection selects against cells expressing the negative selection marker. Thus, cells that have retained the integrated vector express the negative marker and are selectively eliminated. The cells surviving both rounds of selection are those that initially integrated and then eliminated the vector. These cells are enriched for cells having genes modified by homologous recombination with the vector. This process diversifies by a single exchange of genetic information. However, if the process is repeated either with the same vectors or with a library of fragments generated by PCR of pooled DNA from the enriched recombinant population, resulting in the diversity of targeted genes being enhanced exponentially each round of recombination. This process can be repeated recursively, with selection being performed as desired.

#### 8.11.3.4 EXPLOITING KNOWN INFORMATION SUCH AS MAP LOCATION OR FUNCTION

In general, the above methods do not require knowledge of the number of genes to be optimized, their map location or their function. However, in some instances, where this information is available for one or more gene, it can be exploited. For example, if the property to be acquired by evolution is enhanced recombination of cells, one gene likely to be important is *recA*, even though many other genes, known and unknown, may make additional contributions. In this situation, the *recA* gene can be evolved, at least in part, separately from other candidate genes. The *recA* gene can be evolved by any of the methods of recursive recombination described in Section V. Briefly, this approach entails obtaining, diverse forms of a *recA* gene, allowing the forms to recombine, selecting recombinants having improved properties, and subjecting the recombinants to further cycles of recombination and selection. At any point in the individualized improvement of *recA*, the diverse forms of *recA* can be pooled with fragments encoding other genes in a library to be used in the general methods described herein. In this way, the library is seeded to contain a higher proportion of variants in a gene known to be important to the property sought to be acquired than would otherwise be the case.

In one example, a plasmid is constructed carrying a non-functional (mutated) version of a chromosomal gene such as *URA3*, where the wild-type gene confers sensitivity to a drug (in this case 5-fluoro orotic acid). The plasmid also carries a selectable marker (resistance to another drug such as kanamycin), and a library of *recA* variants. Transformation of the plasmid into the cell results in expression of the *recA* variants, some of which will catalyze homologous recombination at an increased rate. Those cells in which homologous recombination occurred are resistant to the selectable drug on the plasmid, and to 5-fluoro orotic acid because of the disruption of the chromosomal copy of this gene.

The *recA* variants which give the highest rates of homologous recombination are the most highly represented in a pool of homologous recombinants. The mutant *recA*



genes can be isolated from this pool by PCR, re-stochastic &/or non-stochastic mutagenized, cloned back into the plasmid and the process repeated. Other sequences can be inserted in place of *recA* to evolve other components of the homologous recombination system.

#### 8.11.3.5 USING OWN HARVEST OF CELLS SO NO IMPURITIES

In some stochastic &/or non-stochastic mutagenesis methods, DNA substrates are isolated from natural sources and are not easily manipulated by DNA modifying or polymerizing enzymes due to recalcitrant impurities, which poison enzymatic reactions. Such difficulties can be avoided by processing DNA substrates through a harvesting strain. The harvesting strain is typically a cell type with natural competence and a capacity for homologous recombination between sequences with substantial diversity (e.g., sequences exhibiting only 75% sequence identity). The harvesting strain bears a vector encoding a negative selection marker flanked by two segments respectively complementary to two segments flanking a gene or other region of interest in the DNA from a target organism. The harvesting strain is contacted with fragments of DNA from the target organism. Fragments are taken up by natural competence, or other methods described herein, and a fragment of interest from the target organism recombines with the vector of the harvesting strain causing loss of the negative selection marker. Selection against the negative marker allows isolation of cells that have taken up the fragment of interest.

Stochastic &/or non-stochastic mutagenesis can be carried out in the harvester strain (e.g., a *RecE/T* strain) or vector can be isolated from the harvester strain for in vitro stochastic &/or non-stochastic mutagenesis or transfer to a different cell type for in vivo stochastic &/or non-stochastic mutagenesis. Alternatively, the vector can be transferred to a different cell type by conjugation, protoplast fusion or electrofusion. An example of a suitable harvester strain is *Acinetobacter calcoaceticus mutS*. Melnikov and Youngman, (1999) *Nucl Acid Res* 27(4):1056-1062. This strain is naturally competent and takes up DNA in a nonsequence-specific manner. Also, because of the *mutS* mutation, this strain is

capable of homologous recombination of sequences showing only 75% sequence identity.

## 8.12 FURTHER APPLICATIONS

### 8.12.1 IMPROVED RECOMBINANCY

One goal of whole cell evolution is to generate cells having improved capacity for recombination. Such cells are useful for a variety of purposes in molecular genetics including the *in vivo* formats of recursive sequence recombination described in Section V. Almost thirty genes (e.g., *recA*, *recB*, *recC*, *recD*, *recE*, *recF*, *recG*, *recO*, *recQ*, *recR*, *recT*, *ruvA*, *ruvB*, *ruvC*, *sbcB*, *ssb*, *topA*, *gyrA* and *B*, *lig*, *polA*, *uvrD*, *E*, *recL*, *mutD*, *mutH*, *mutL*, *mutT*, *mutU*, *helD*) and DNA sites (e.g., *chi*, *recN*, *sbcC*) involved in genetic recombination have been identified in *E. coli*, and cognate forms of several of these genes have been found in other organisms (e.g., *rad51*, *rad55-rad57*, *Dmcl* in yeast (see Kowalczykowski et al., *Microbiol. Rev.* 58, 401-465 (1994); Kowalczykowski & Zarling, *supra*) and human homologs of *Rad51* and *Dmcl* have been identified (see Sandier et al., *Nucl. Acids Res.* 24, 2125-2132 (1996)). At least some of the *E. coli* genes, including *recA* are functional in mammalian cells, and can be targeted to the nucleus as a fusion with SV40 large T antigen nuclear targeting sequence (Reiss et al., *Proc. Mad. Acad. Sci. USA*, 93, 3094-3098 (1996)). Further, mutations in mismatch repair genes, such as *mutL*, *mutS*, *mutH* *mutT* relax homology requirements and allow recombination between more diverged sequences (Rayssiguier et al., *Nature* 342, 396-401 (1989)). The extent of recombination between divergent strains can be enhanced by impairing mismatch repair genes and stimulating SOS genes. Such can be achieved by use of appropriate mutant strains and/or growth under conditions of metabolic stress, which have been found to stimulate SOS and inhibit mismatch repair genes. Vulic et al., *Proc. Mad. Acad. Sci. USA* 94 (1997). In addition, this can be achieved by impairing the products of mismatch repair genes by exposure to selective inhibitors.

Starting substrates for recombination are selected according to the general principles described above. That is, the substrates can be whole genomes or fractions

thereof containing recombination genes or sites. Large libraries of essentially random fragments can be seeded with collections of fragments constituting variants of one or more known recombination genes, such as *recA*. Alternatively, libraries can be formed by mixing variant forms of the various known recombination genes and sites.

#### **8.12.2 EXPRESSION OF *GFP* INDICATES CELL IS CAPABLE OF HOMOLOGOUS RECOMBINATION**

The library of fragments is introduced into the recipient cells to be improved and recombination occurs, generating modified cells. The recipient cells preferably contain a marker gene whose expression has been disabled in a manner that can be corrected by recombination. For example, the cells can contain two copies of a marker gene bearing mutations at different sites, which copies can recombine to generate the wild type gene. A suitable marker gene is green fluorescent protein. A vector can be constructed encoding one copy of GFP having stop codons near the N-terminus, and another copy of GFP having stop codons near the C-terminus of the protein. The distance between the stop codons at the respective ends of the molecule is 500 bp and about 25% of recombination events result in active GFP. Expression of GFP in a cell signals that a cell is capable of homologous recombination to recombine in between the stop codons to generate a contiguous coding sequence. By screening for cells expressing GFP, one enriches for cells having the highest capacity for recombination. The same type of screen can be used following subsequent rounds of recombination. However, unless the selection marker used in previous round(s) was present on a suicide vector, subsequent round(s) should employ a second disabled screening marker within a second vector bearing a different origin of replication or a different positive selection marker to vectors used in the previous rounds.

#### **8.12.3 INCREASED GENOME COPY NUMBER SO MORE CHROMOSOMES PER BACTERIAL CELL TO MAKE EVOLUTION QUICKER**

The majority of bacterial cells in stationary phase cultures grown in rich media contain two, four or eight genomes. In minimal medium the cells contain one or two

genomes. The number of genomes per bacterial cell thus depends on the growth rate of the cell as it enters stationary phase. This is because rapidly growing cells contain multiple replication forks, resulting in several genomes in the cells after termination. The number of genomes is strain dependent, although all strains tested have more than one chromosome in stationary phase. The number of genomes in stationary phase cells decreases with time. This appears to be due to fragmentation and degradation of entire chromosomes, similar to apoptosis in mammalian cells. This fragmentation of genomes in cells containing multiple genome copies results in massive recombination and mutagenesis. Useful mutants may find ways to use energy sources that will allow them to continue growing. Multigenome or gene-redundant cells are much more resistant to mutagenesis and can be improved for a selected trait faster.

Some cell types, such as *Deinococcus radians* (Daly and Minton *J Bacteriol* 177, 5495-5505 (1995)) exhibit polyploidy throughout the cell cycle. This cell type is highly radiation resistant due to the presence of many copies of the genome. High frequency recombination between the genomes allows rapid removal of mutations induced by a variety of DNA damaging agents.

A goal of the present methods is to evolve other cell types to have increased genome copy number akin to that of *Deinococcus radians*. Preferably, the increased copy number is maintained through all or most of its cell cycle in all or most growth conditions. The presence of multiple genome copies in such cells results in a higher frequency of homologous recombination in these cells, both between copies of a gene in different genomes within the cell, and between a genome within the cell and a transfected fragment. The increased frequency of recombination allows the cells to be evolved more quickly to acquire other useful characteristics.

Starting substrates for recombination can be a diverse library of genes only a few of which are relevant to genomic copy number, a focused library formed from variants of gene(s) known or suspected to have a role in genomic copy number or a combination of the two. As a general rule one would expect increased copy number would be achieved by

evolution of genes involved in replication and cell septation such that cell septation is inhibited without impairing replication. Genes involved in replication include *tus*, *xerC*, *xerD*, *dif*, *gyrA*, *gyrB*, *parE*, *parC*, *dif*, *TerA*, *TerB*, *TerC*, *TerD*, *TerE*, *TerF*, and genes influencing chromosome partitioning and gene copy number include *minD*, *mukA* (*tolC*), *mukB*, *mukC*, *mukD*, *spoOJ*, *spoIIIE* (Wake & Errington, *Annu. Rev. Genet.* 29, 41-67 (1995)). A useful source of substrates is the genome of a cell type such as *Deinococcus radians* known to have the desired phenotype of multigenomic copy number. As well as, or instead of, the above substrates, fragments encoding protein or antisense RNA inhibitors to genes known to be involved in cell septation can also be used. In nature, the existence of multiple genomic copies in a cell type would usually not be advantageous due to the greater nutritional requirements needed to maintain this copy number. However, artificial conditions can be devised to select for high copy number.

Modified cells having recombinant genomes are grown in rich media (in which conditions, multicopy number should not be a disadvantage) and exposed to a mutagen, such as ultraviolet or gamma irradiation or a chemical mutagen, e.g., mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, which induces DNA breaks amenable to repair by recombination. These conditions select for cells having multicopy number due to the greater efficiency with which mutations can be excised. Modified cells surviving exposure to mutagen are enriched for cells with multiple genome copies. If desired, selected cells can be individually analyzed for genome copy number (e.g., by quantitative hybridization with appropriate controls). Some or all of the collection of cells surviving selection provide the substrates for the next round of recombination. In addition, individual cells can be sorted using a cell sorter for those cells containing more DNA, e.g., using DNA specific fluorescent compounds or sorting for increased size using light dispersion. Eventually cells are evolved that have at least 2, 4, 6, 8 or 10 copies of the genome throughout the cell cycle. In a similar manner, protoplasts can also be recombined.

#### **8.12.4 EVOLVE SECRETION PATHWAYS FOR BETTER EFFICIENCY**

#### **8.12.5 EVOLVE TO MANUFACTURE DRUGS OR CHEMICALS**

The protein (or metabolite) secretion pathways of bacterial and eukaryotic cells can be evolved to export desired molecules more efficiently, such as for the manufacturing of protein pharmaceuticals, small molecule drugs or specialty chemicals. Improvements in efficiency are particularly desirable for proteins requiring multisubunit assembly (such as antibodies) or extensive posttranslational modification before secretion.

The efficiency of secretion may depend on a number of genetic sequences including a signal peptide coding sequence, sequences encoding protein(s) that cleave or otherwise recognize the coding sequence, and the coding sequence of the protein being secreted. The latter may affect folding of the protein and the ease with which it can integrate into and traverse membranes. The bacterial secretion pathway in *E. coli* include the SecA, SecB, SecE, SecD and SecF genes. In *Bacillus subtilis*, the major genes are secA, secD, secE, secF, secY, ffh, ftsY together with five signal peptidase genes (sipS, sipT, sipU, sipV and sipW) (Kunst et al, supra). For proteins requiring posttranslational modification, evolution of genes effecting such modification may contribute to improved secretion. Likewise genes with expression products having a role in assembly of multisubunit proteins (e.g., chaperonins) may also contribute to improved secretion.

Selection of substrates for recombination follows the general principles discussed above. In this case, the focused libraries referred to above comprise variants of the known secretion genes. For evolution of prokaryotic cells to express eukaryotic proteins, the initial substrates for recombination are often obtained at least in part from eukaryotic sources.

Incoming fragments can undergo recombination both with chromosomal DNA in recipient cells and with the screening marker construct present in such cells (see below). The latter form of recombination is important for evolution of the signal coding sequence incorporated in the screening marker construct. Improved secretion can be screened by the inclusion of marker construct in the cells being evolved. The marker construct encodes a

marker gene, operably linked to expression sequences, and usually operably linked to a signal peptide coding sequence. The marker gene is sometimes expressed as a fusion protein with a recombinant protein of interest. This approach is useful when one wants to evolve the recombinant protein coding sequence together with secretion genes.

#### 8.12.6 EVOLVE SO PRODUCT IS TOXIC TO CELL UNLESS SECRETED

In one variation, the marker gene encodes a product that is toxic to the cell containing the construct unless the product is secreted. Suitable toxin proteins include diphtheria toxin and ricin toxin. Propagation of modified cells bearing such a construct selects for cells that have evolved to improve secretion of the toxin. Alternatively, the marker gene can encode a ligand to a known receptor, and cells bearing the ligand can be detected by FACS using labeled receptor. Optionally, such a ligand can be operably linked to a phospholipid anchoring sequence that binds the ligand to the cell membrane surface following secretion. In a further variation, secreted marker protein can be maintained in proximity with the cell secreting it by distributing individual cells into agar drops. This is done, e.g., by droplet formation of a cell suspension. Secreted protein is confined within the agar matrix and can be detected by e.g., FACS. In another variation, a protein of interest is expressed as a fusion protein together with beta-lactamase or alkaline phosphatase. These enzymes metabolize commercially available chromogenic substrates (e.g., X-gal), but do so only after secretion into the periplasm. Appearance of colored substrate in a colony of cells therefore indicates capacity to secrete the fusion protein and the intensity of color is related to the efficiency of secretion.

The cells identified by these screening and selection methods have the capacity to secrete increased amounts of protein. This capacity may be attributable to increased secretion and increased expression, or from increased secretion alone.

### **8.12.7 EVOLVE TO ACQUIRE INCREASED EXPRESSION OF RECOMBINANT PROTEIN**

Expression Cells can also be evolved to acquire increased expression of a recombinant protein. The level of expression is, of course, highly dependent on the construct from which the recombinant protein is expressed and the regulatory sequences, such as the promoter, enhancer(s) and transcription termination site contained therein. Expression can also be affected by a large number of host genes having roles in transcription, posttranslational modification and translation. In addition, host genes involved in synthesis of ribonucleotide and amino acid monomers for transcription and translation may have indirect effects on efficiency of expression. Selection of substrates for recombination follows the general principles discussed above. In this case, focused libraries comprise variants of genes known to have roles in expression. For evolution of prokaryotic cells to express eukaryotic proteins, the initial substrates for recombination are often obtained, at least in part, from eukaryotic sources; that is eukaryotic genes encoding proteins such as chaperonins involved in secretion and/assembly of proteins. Incoming fragments can undergo recombination both with chromosomal DNA in recipient cells and with the screening marker construct present in such cells (see below).

Screening for improved expression can be effected by including a reporter construct in the cells being evolved. The reporter construct expresses (and usually secretes) a reporter protein, such as GFP, which is easily detected and nontoxic. The reporter protein can be expressed alone or together with a protein of interest as a fusion protein. If the reporter gene is secreted, the screening effectively selects for cells having either improved secretion or improved expression, or both.

### **8.12.8 EVOVLE PLANT CELLS TO ACQUIRE RESISTANCE**

A further application of recursive sequence recombination is the evolution of plant cells, and transgenic plants derived from the same, to acquire resistance to pathogenic diseases (fungi, viruses and bacteria), insects, chemicals (such as salt, selenium, pollutants, pesticides, herbicides, or the like), including, e.g., atrazine or glyphosate, or to modify



chemical composition, yield or the like. The substrates for recombination can again be whole genomic libraries, fractions thereof or focused libraries containing variants of gene(s) known or suspected to confer resistance to one of the above agents. Frequently, library fragments are obtained from a different species to the plant being evolved.

The DNA fragments are introduced into plant tissues, cultured plant cells, plant microspores, or plant protoplasts by standard methods including electroporation (From et al., Proc. Natl. Acad. Sci. USA 82, 5824 (1985), infection by viral vectors such as cauliflower mosaic virus (CaMV) (Hohn et al., Molecular Biology of Plant Tumors, (Academic Press, New York, 1982) pp. 549-560; Howell, US 4,407,956), high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface (Klein et al., Nature 327, 70-73 (1987)), use of pollen as vector (WO 85/01856), or use of *Agrobacterium tumefaciens* or *A. rhizogenes* carrying a T-DNA plasmid in which DNA fragments are cloned. The T-DNA plasmid is transmitted to plant cells upon infection by *Agrobacterium tumefaciens*, and a portion is stably integrated into the plant genome (Horsch et al., Science 233, 496-498 (1984); Fraley et al., Proc. Natl. Acad. Sci. USA 80, 4803 (1983)).

Diversity can also be generated by genetic exchange between plant protoplasts according to the same principles described below for fungal protoplasts. Procedures for formation and fusion of plant protoplasts are described by Takahashi et al., US 4,677,066; Akagi et al., US 5,360,725; Shimamoto et al., US 5,250,433; Cheney et al., US 5,426,040.

#### **8.12.9 PLANT GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

Plant genome stochastic &/or non-stochastic mutagenesis allows recursive cycles to be used for the introduction and recombination of genes or pathways that confer improved properties to desired plant species. Any plant species, including weeds and wild cultivars, showing a desired trait, such as herbicide resistance, salt tolerance, pest resistance, or temperature tolerance, can be used as the source of DNA that is introduced into the crop or horticultural host plant species.

Genomic DNA prepared from the source plant is fragmented (e.g. by DNaseI, restriction enzymes, or mechanically) and cloned into a vector suitable for making plant genomic libraries, such as pGA482 (An. G., 1995, Methods Mol. Biol. 44:47-58). This vector contains the *A. tumefaciens* left and right borders needed for gene transfer to plant cells and antibiotic markers for selection in *E. coli*, *Agrobacterium*, and plant cells. A multicloning site is provided for insertion of the genomic fragments. A cos sequence is present for the efficient packaging of DNA into bacteriophage lambda heads for transfection of the primary library into *E. coli*. The vector accepts DNA fragments of 25-40 kb.

The primary library can also be directly electroporated into an *A. tumefaciens* or *A. rhizogenes* strain that is used to infect and transform host plant cells (Main, GD et al., 1995, Methods Mol. Biol. 44:405-412). Alternatively, DNA can be introduced by electroporation or PEG-mediated uptake into protoplasts of the recipient plant species (Bilang et al. (1994) Plant Mol. Biol Manual, Kluwer Academic Publishers, Al: 1- 16) or by particle bombardment of cells or tissues (Christou, *ibid*, A2:1-15). If necessary, antibiotic markers in the T-DNA region can be eliminated, as long as selection for the trait is possible, so that the final plant products contain no antibiotic genes.

Stably transformed whole cells acquiring the trait are selected on solid or liquid media containing the agent to which the introduced DNA confers resistance or tolerance. If the trait in question cannot be selected for directly, transformed cells can be selected with antibiotics and allowed to form callus or regenerated to whole plants and then screened for the desired property.

The second and further cycles consist of isolating genomic DNA from each transgenic line and introducing it into one or more of the other transgenic lines. In each round, transformed cells are selected or screened for incremental improvement. To speed the process of using multiple cycles of transformation, plant regeneration can be deferred until the last round. Callus tissue generated from the protoplasts or transformed tissues can

serve as a source of genomic DNA and new host cells. After the final round, fertile plants are regenerated and the progeny are selected for homozygosity of the inserted DNAs. Ultimately, a new plant is created that carries multiple inserts which additively or synergistically combine to confer high levels of the desired trait. Alternatively, microspores can be isolated as homozygotes generated from spontaneous diploids.

In addition, the introduced DNA that confers the desired trait can be traced because it is flanked by known sequences in the vector. Either PCR or plasmid rescue is used to isolate the sequences and characterize them in more detail. Long PCR (Foord, OS and Rose, EA, 1995, PCR Primer: A Laboratory Manual, CSBL Press, pp 63-77) of the full 25-40 kb insert is achieved with the proper reagents and techniques using as primers the T-DNA border sequences. If the vector is modified to contain the *E. coli* origin of replication and an antibiotic marker between the T-DNA borders, a rare cutting restriction enzyme, such as NotI or SfiI, that cuts only at the ends of the inserted DNA is used to create fragments containing the source plant DNA that are then self-ligated and transformed into *E. coli* where they replicate as plasmids. The total DNA or subfragment of it that is responsible for the transferred trait can be subjected to in vitro evolution by DNA stochastic &/or non-stochastic mutagenesis. The stochastic &/or non-stochastic mutagenized library can be reiteratively recombined by any method herein and then introduced into host plant cells and screened for improvement of the trait. In this way, single and multigene traits can be transferred from one species to another and optimized for higher expression or activity leading to whole organism improvement. This entire process can also be reiteratively repeated. Alternatively, the cells can be transformed microspores with the regenerated haploid plants being screened directly for improved traits as noted below.

#### **8.12.10 PLANT CELL IS PUT INTO CONTACT WITH AGENT TO SEE WHICH CELLS SURVIVE.**

After a suitable period of incubation to allow recombination to occur and for expression of recombinant genes, the plant cells are contacted with the agent to which

resistance is to be acquired, and surviving plant cells are collected. Some or all of these plant cells can be subject to a further round of recombination and screening. Eventually, plant cells having the required degree of resistance are obtained.

These cells can then be cultured into transgenic plants. Plant regeneration from cultured protoplasts is described in Evans et al., "Protoplast Isolation and Culture," Handbook of Plant Cell Cultures 1, 124-176 (MacMillan Publishing Co., New York, 1983); Davey, "Recent Developments in the Culture and Regeneration of Plant Protoplasts," Protoplasts, (1983) pp. 12-29, (Birkhauser, Basel 1983); Dale, "Protoplast Culture and Plant Regeneration of Cereals and Other Recalcitrant Crops," Protoplasts (1983) pp. 31-41, (Birkhauser, Basel 1983); Binding, "Regeneration of Plants," Plant Protoplasts, pp. 21-73, (CRC Press, Boca Raton, 1985).

#### **8.12.11 START IN BACTERIAL CELL SINCE FASTER EVOLUTION AND TRANSFORM INTO PLANT**

In a variation of the above method, one or more preliminary rounds of recombination and screening can be performed in bacterial cells according to the same general strategy as described for plant cells. More rapid evolution can be achieved in bacterial cells due to their greater growth rate and the greater efficiency with which DNA can be introduced into such cells. After one or more rounds of recombination/screening, a DNA fragment library is recovered from bacteria and transformed into the plant cells. The library can either be a complete library or a focused library. A focused library can be produced by amplification from primers specific for plant sequences, particularly plant sequences known or suspected to have a role in conferring resistance.

#### **8.12.12 MICROSPORE MANIPULATION**

Microspores are haploid (1n) male spores that develop into pollen grains. Anthers contain a large numbers of microspores in early-uninucleate to first-mitosis stages. Microspores have been successfully induced to develop into plants for most species, such

as, e.g., rice (Chen, CC 1977 *In Vitro*. 13: 484-489), tobacco (Atanasov, I. et al. 1998 *Plant Mol Biol*. 38:1169-1178), *Tradescantia* (Savage JRK and Papworth DG. 1998 *Mutat Res*. 422:313-322), *Arabidopsis* (Park SK et al. 1998 *Development*. 125:3789- 3799), sugar beet (Majewska-Sawka A and Rodrigues-Garcia NE 1996 *J Cell Sci*. 109:859-866), Barley (Olsen FL 1991 *Hereditas* 115:255-266) and oilseed rape (Boutillier KA et al. 1994 *Plant Mol Biol*. 26:1711-1723).

The plants derived from microspores are predominantly haploid or diploid (infrequently polyploid and aneuploid). The diploid plants are homozygous and fertile and can be generated in a relatively short time. Microspores obtained from F1 hybrid plants represent great diversity, thus being an excellent model for studying recombination. In addition, microspores can be transformed with T-DNA introduced by agrobacterium or other available means and then regenerated into individual plants. Furthermore, protoplasts can be made from microspores and they can be fused similar to what occur in fungi and bacteria.

Microspores, due to their complex ploidy and regenerating ability, provide a tool for plant whole genome stochastic &/or non-stochastic mutagenesis. For example, if pollens from 4 parents are collected 4 and pooled, and then used to randomly pollinate the parents, the progenies should have  $2^4 = 16$  possible combinations. Assuming this plant has 7 chromosomes, microspores collected from the 16 progenies will represent  $2^7 \times 16 = 2048$  possible chromosomal combinations. This number is even greater if meiotic processes occur. When diploid, homozygous embryos are generated from these microspores, in many cases, they are screened for desired phenotypes, such as herbicide- or disease- resistant. In addition, for plant oil composition these embryos can be dissected into two halves: one for analysis the other for regeneration into a viable plant. Protoplasts generated from microspores (especially the haploid ones) are pooled and fused. Microspores obtained from plants generated by protoplast fusion are pooled and fused again, increasing the genetic diversity of the resulting microspores. Microspores can be subjected to mutagenesis in various ways, such as by chemical mutagenesis, radiation-induced mutagenesis and, e.g., t-DNA transformation, prior to fusion or regeneration. New

mutations which are generated can be recombined through the recursive processes described above and herein.

### 8.12.13 ACQUISITION OF SALT TOLERANCE

DNA from a salt tolerant plant is isolated and used to create a genomic library. Protoplasts made from the recipient species are transformed/transfected with the genomic library (e.g., by electroporation, agrobacterium, etc.). Cells are selected on media with a normally inhibitory level of NaCl. Only the cells with newly acquired salt tolerance will grow into callus tissue. The best lines are chosen and genomic libraries are made from their pooled DNA. These libraries are transformed into protoplasts made from the first round transformed calli. Again, cells are selected on increased salt concentrations. After the desired level of salt tolerance is achieved, the callus tissue can be induced to regenerate whole plants. Progeny of these plants are typically analyzed for homozygosity of the inserts to ensure stability of the acquired trait. At the indicated steps, plant regeneration or isolation and stochastic &/or non-stochastic mutagenesis of the introduced genes can be added to the overall protocol.

### 8.13 EVOLVE TRANSGENIC ANIMALS

#### 8.13.1 OPTIMIZE TRANSGENE

One goal of transgenesis is to produce transgenic animals, such as mice, rabbits, sheep, pigs, goats, and cattle, secreting a recombinant protein in the milk. A transgene for this purpose typically comprises in operable linkage a promoter and an enhancer from a milk-protein gene (e.g., alpha, beta, or gamma casein, beta-lactoglobulin, acid whey protein or alpha-lactalbumin), a signal sequence, a recombinant protein coding sequence and a transcription termination site.

Optionally, a transgene can encode multiple chains of a multichain protein, such as an immunoglobulin, in which case, the two chains are usually individually operably linked to sets of regulatory sequences. Transgenes can be optimized for expression and secretion

by recursive sequence recombination. Suitable substrates for recombination include regulatory sequences such as promoters and enhancers from milk-protein genes from different species or individual animals. Cycles of recombination can be performed in vitro or in vivo by any of the formats discussed. Screening is performed in vivo on cultures of mammary-gland derived cells, such as HC11 or MacT, transfected with transgenes and reporter constructs such as those discussed above. After several cycles of recombination and screening, transgenes resulting in the highest levels of expression and secretion are extracted from the mammary gland tissue culture cells and used to transfect embryonic cells, such as zygotes and embryonic stem cells, which are matured into transgenic animals.

### **8.13.2 OPTIMIZE WHOLE ANIMAL BY TRANSFORMING INTO EMBRYONIC CELLS GENE OF DESIRED TRAIT**

#### **8.13.2.1 GROWTH HORMONE**

In this approach, libraries of incoming fragments are transformed into embryonic cells, such as ES cells or zygotes. The fragments can be variants of a gene known to confer a desired property, such as growth hormone. Alternatively, the fragments can be partial or complete genomic libraries including many genes. Fragments are usually introduced into zygotes by microinjection as described by Gordon et al., *Methods Enzymol.* 10 1, 414 (1984); Hogan et al., *Manipulation of the Mouse Embryo: A Laboratory Manual* (C.S.H.L. N.Y., 1986) (mouse embryo),- and Hammer et al., *Nature* 315, 680 (1985) (rabbit and porcine embryos); Gandolfi et al., *J Reprod. Fert.* 81, 23-28 (1987); Rexroad et al., *J Anim. Sci.* 66, 947-953 (1988) (ovine embryos) and Eyestone et al., *J Reprod. Fert.* 85, 715-720 (1989); Camous et al., *J Reprod. Fert.* 72, 779- 785 (1984); and Heyman et al., *Theriogenology* 27, 5968 (1987) (bovine embryos). Zygotes are then matured and introduced into recipient female animals which gestate the embryo and give birth to a transgenic offspring.

Alternatively, transgenes can be introduced into embryonic stem cells (ES).

These cells are obtained from preimplantation embryos cultured in vitro. Bradley et al., *Nature* 309, 255-258 (1984). Transgenes can be introduced into such cells by

electroporation or microinjection. Transformed ES cells are combined with blastocysts from a non-human animal. The ES cells colonize the embryo and in some embryos form the germ line of the resulting chimeric animal. See Jaenisch, Science, 240, 1468-1474 (1988).

Regardless whether zygotes or ES are used, screening is performed on whole animals for a desired property, such as increased size and/or growth rate. DNA is extracted from animals having evolved toward acquisition of the desired property. This DNA is then used to transfect further embryonic cells. These cells can also be obtained from animals that have acquired toward the desired property in a split and pool approach. That is, DNA from one subset of such animals is transformed into embryonic cells prepared from another subset of the animals. Alternatively, the DNA from animals that have evolved toward acquisition of the desired property can be transfected into fresh embryonic cells. In either alternative, transfected cells are matured into transgenic animals, and the animals subjected to a further round of screening for the desired property.

Initially, a library is prepared of variants of a growth hormone gene. The variants can be natural or induced. The library is coated with recA protein and transfected into fertilized fish eggs. The fish eggs then mature into fish of different sizes. The growth hormone gene fragment of genomic DNA from large fish is then amplified by PCR and used in the next round of recombination. Alternatively, fish -IFN is evolved to enhance resistance to viral infections as described below.

#### **8.13.2.2 EVOLUTION OF IMPROVED HORMONES FOR EXPRESSION IN TRANSGENIC ANIMALS**

#### **8.13.3 TO CREATE ANIMALS WITH IMPROVED TRAITS**

Evolution of improved hormones for expression in transgenic animals (e.g., Fish) to create animals with improved traits. Hormones and cytokines are key regulators of size, body weight, viral resistance and many other commercially important traits. DNA stochastic &/or non-stochastic mutagenesis is used to rapidly evolve the genes for these



proteins using in vitro assays. This was demonstrated with the evolution of the human alpha interferon genes to have potent antiviral activity on murine cells. Large improvements in activity were achieved in two cycles of family stochastic &/or non-stochastic mutagenesis of the human IFN genes.

In general, a method of increasing resistance to virus infection in cells can be performed by first introducing a stochastic &/or non-stochastic mutagenized library comprising at least one stochastic &/or non-stochastic mutagenized interferon gene into animal cells to create an initial library of animal cells or animals. The initial library is then challenged with the virus. Animal cells or animals are selected from the initial library which are resistant to the virus and a plurality of transgenes from a plurality of animal cells or animals which are resistant to the virus are recovered. The plurality of transgenes is recovered to produce an evolved library of animal cells or animals which is again challenged with the virus. Cells or animals are selected from the evolved library the which are resistant to the virus.

For example, genes evolved with in vitro assays are introduced into the germplasm of animals or plants to create improved strains. One limitation of this procedure is that in vitro assays are often only crude predictors of in vivo activity. However, with improving methods for the production of transgenic plants and animals, one can now marry whole organism breeding with molecular breeding. The approach is to introduce stochastic &/or non-stochastic mutagenized libraries of hormone genes into the species of interest. This can be done with a single gene per transgenic or with pools of genes per transgenic. Progeny are then screened for the phenotype of interest. In this case, stochastic &/or non-stochastic mutagenized libraries of interferon genes (alpha IFN for example) are introduced into transgenic fish. The library of transgenic fish are challenged with a virus. The most resistant fish are identified (i.e. either survivors of a lethal challenge; or those that are deemed most 'healthy' after the challenge). The IFN transgenes are recovered by PCR and stochastic &/or non-stochastic mutagenized in either a poolwise or a pairwise fashion. This generates an evolved library of IFN genes. A second library of transgenic fish is created and the process is repeated. In this way, IFN is evolved for improved

antiviral activity in a whole organism assay. This procedure is general and can be applied to any trait that is affected by a gene or gene family of interest and which can be quantitatively measured.

Fish interferon sequence data is available for the Japanese flatfish (*Paralichthys olivaceus*) as mRNA sequence (Tamai et al (1993) "Cloning and expression of flatfish (*Paralichthys olivaceus*) interferon cDNA." *Biochem. Biophys. Acta* 1174, 182-186; Y see also, Tami et al. (1993) "Purification and characterization of interferon-like antiviral protein derived from flatfish (*Paralichthys olivaceus*) lymphocytes immortalized by oncogenes." *Cytotechnology* 1993; 1 1 (2):121-131). This sequence can be used to clone out IFN genes from this species. This sequence can also be used as a probe to clone homologous interferons from additional species of fish. As well, additional sequence information can be utilized to clone out more species of fish interferons. Once a library of interferons has been cloned, these can be family stochastic &/or non-stochastic mutagenized to generate a library of variants.

In one embodiment, BHK-21 (A fibroblast cell line from hamster) can be transfected with the stochastic &/or non-stochastic mutagenized WN-expression plasmids. Active recombinant IFN is produced and then purified by WGA agarose affinity chromatography (Tamai, et al. 1993 *Biochim Biophys Acta. supra*). The antiviral activity of IFN can be measured on fish cells challenged by rhabdovirus. Tami et al. (1993) "Purification and characterization of interferon-like antiviral protein derived from flatfish (*Paralichthys olivaceus*) lymphocytes immortalized by oncogenes. " *Cytotechnology* 1993; 1 1 (2):121-131).

#### **8.13.4 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS IN HIGHER ORGANISMS**

**POOLWISE RECURSIVE BREEDING** The present invention provides a procedure for generating large combinatorial libraries of higher eukaryotes, plants, fish, domesticated animals, etc. In addition to the procedures outlined above, poolwise

combination of male and female gametes can also be used to generate large diverse molecular libraries.

In one aspect, the process includes recursive poolwise matings for several generations without any deliberate screening. This is similar to classical breeding, except that pools of organisms, rather than pairs of organisms, are mated, thereby accelerating the generation of genetic diversity. This method is similar to recursive fusion of a diverse population of bacterial protoplasts resulting in the generation of multiparent progeny harboring genetic information from all of the starting population of bacteria. The process described here is to perform analogous artificial or natural matings of large populations of natural isolates, imparting a split pool mating strategy. Before mating, all of the male gametes i.e. pollen, sperm, etc., are isolated from the starting population and pooled. These are then used to "self" fertilize a mixed pool of the female gametes from the same population.

The process is repeated with the subsequent progeny for several generations, with the final progeny being a combinatorial organism library with each member having genetic information originating from many if not all of the starting "parents." This process generates large diverse organism libraries on which many selections and or screens can be imparted, and it does not require sophisticated in vitro manipulation of genes. However, it results in the creation of useful new strains (perhaps well diluted in the population) in a much shorter time frame than such organisms could be generated using a classical targeted breeding approach.

These libraries are generated relatively quickly (e.g., typically in less than three years for most plants of commercial interest, with six cycles or less of recursive breeding being sufficient to generate desired diversity). An additional benefit of these methods is that the resulting libraries provide organismal diversity in areas, such as agriculture, aquaculture, and animal husbandry, that are currently genetically homogeneous.

Examples of these methods for several organisms are described below.

### 8.13.5 PLANTS

**Plants** A population of plants, for example all of the different corn strains in a commercial seed/germplasm collection, are grown and the pollen from the entire population is harvested and pooled. This mixed pollen population is then used to "self" fertilize the same population. Self pollination is prevented, so that the fertilization is combinatorial. The cross results in all pairwise crosses possible within the population, and the resulting seeds result in many of the possible outcomes of each of these pairwise crosses. The seeds from the fertilized plants are then harvested, pooled, planted, and the pollen is again harvested, pooled, and used to "self" fertilize the population. After only several generations, the resulting population is a very diverse combinatorial library of corn. The seeds from this library are harvested and screened for desirable traits, e.g., salt tolerance, growth rate, productivity, yield, disease resistance, etc. Essentially any plant collection can be modified by this approach. Important commercial crops include both monocots and dicots. Monocots include plants in the grass family (Gramineae), such as plants in the sub families Fetucoeidae and Poacoideae, which together include several hundred genera including plants in the genera *Agrostis*, *Phleum*, *Dactylis*, *Sorgum*, *Setaria*, *Zea* (e.g., corn), *Oryza* (e.g., rice), *Triticum* (e.g., wheat), *Secale* (e.g., rye), *Avena* (e.g., oats), *Hordeum* (e.g., barley), *Saccharum*, *Poa*, *Festuca*, *Stenotaphrum*, *Cynodon*, *Coix*, the *Olyreae*, *Phareae* and many others. Plants in the family Gramineae are a particularly preferred target plants for the methods of the invention.

Additional preferred targets include other commercially important crops, e.g., from the families *Compositae* (the largest family of vascular plants, including at least 1,000 genera, including important commercial crops such as sunflower), and *Leguminosae* or "pea family," which includes several hundred genera, including many commercially valuable crops such as pea, beans, lentil, peanut, yam bean, cowpeas, velvet beans, soybean, clover, alfalfa, lupine, vetch, lotus, sweet clover, wisteria, and sweetpea. Common crops applicable to the methods of the invention include *Zea mays*, rice, soybean, sorghum, wheat, oats, barley, millet, sunflower, and canola.

This process can also be carried out using pollen from different species or more divergent strains (e.g., crossing the ancient grasses with corn). Different plant species can

be forced to cross. Only a few plants from an initial cross would have to result in order to make the process viable. These few progeny, e.g., from a cross between soy bean and corn, would generate pollen and eggs, each of which would represent a different meiotic outcome from the recombination of the two genomes. The pollen would be harvested and used to "self pollinate the original progeny. This process would then be carried out recursively. This would generate a large family stochastic &/or non-stochastic mutagenized library of two or more species, which could be subsequently screened.

#### **8.13.6 FISH**

**Fish** The natural tendency of fish to lay their eggs outside of the body and to have a male cover those eggs with sperm provides another opportunity for a split pooled breeding strategy. The eggs from many different fish, e.g., salmon from different fisheries about the world, can be harvested, pooled, and then fertilized with similarly collected and pooled salmon sperm. The fertilization will result in all of the possible pairwise matings of the starting population. The resulting progeny is then grown and again the sperm and eggs are harvested, and pooled, with each egg and sperm representing a different meiotic outcome of the different crosses. The pooled sperm are then used to fertilize the pooled eggs and the process is carried out recursively. After several generations the resulting progeny can then be subjected to selections and screens for desired properties, such as size, disease resistance, etc.

#### **8.13.7 ANIMALS**

**Animals** The advent of in vitro fertilization and surrogate motherhood provides a means of whole genome stochastic &/or non-stochastic mutagenesis in animals such as mammals. As with fish, the eggs and the sperm from a population, for example from all slaughter cows, are collected and pooled. The pooled eggs are then in vitro fertilized with the pooled sperm. The resulting embryos are then returned to surrogate mothers for development. As above, this process is repeated recursively until a large diverse population is generated that can be screened for desirable traits.

A technically feasible approach would be similar to that used for plants. In this

case, sperm from the males of the starting population is collected and pooled, and then this pooled sample is used to artificially inseminate multiple females from each of the starting populations. Only one (or a few) sperm would succeed in each animal, but these should be different for each fertilization. The process is reiterated by harvesting the sperm from all of the male progeny, pooling it, and using it to fertilize all of the female progeny. The process is carried out recursively for several generations to generate the organism library, which can then be screened.

#### 8.14 PREDICTIVE TOOL IN LOOKING FOR DRUGS

Recursive sequence recombination can be used to simulate natural evolution of pathogenic microorganisms in response to exposure to a drug under test. Using recursive sequence recombination, evolution proceeds at a faster rate than in natural evolution. One measure of the rate of evolution is the number of cycles of recombination and screening required until the microorganism acquires a defined level of resistance to the drug. The information from this analysis is of value in comparing the relative merits of different drugs and in particular, in predicting their long term efficacy on repeated administration.

The pathogenic microorganisms used in this analysis include the bacteria that are a common source of human infections, such as chlamydia, rickettsial bacteria, mycobacteria, staphylococci, streptococci, pneumonococci, meningococci and gonococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lyme disease bacteria.

Evolution is effected by transforming an isolate of bacteria that is sensitive to a drug under test with a library of DNA fragments. The fragments can be a mutated version of the genome of the bacteria being evolved. If the target of the drug is a known protein or nucleic acid, a focused library containing variants of the corresponding gene can be used. Alternatively, the library can come from other kinds of bacteria, especially bacteria typically found inhabiting human tissues, thereby simulating the source material available for recombination in vivo. The library can also come from bacteria known to be resistant

to the drug. After transformation and propagation of bacteria for an appropriate period to allow for recombination to occur and recombinant genes to be expressed, the bacteria are screened by exposing them to the drug under test and then collecting survivors. Surviving bacteria are subject to further rounds of recombination. The subsequent round can be effected by a split and pool approach in which DNA from one subset of surviving bacteria is introduced into a second subset of bacteria. Alternatively, a fresh library of DNA fragments can be introduced into surviving bacteria. Subsequent round(s) of selection can be performed at increasing concentrations of drug, thereby increasing the stringency of selection.

#### 8.14.1 BIOSYNTHESIS

Metabolic engineering can be used to alter organisms to optimize the production of practically any metabolic intermediate, including antibiotics, vitamins, amino acids such as phenylalanine and aromatic amino acids, ethanol, butanol, polymers such as xanthan gum and bacterial cellulose, peptides, and lipids. When such compounds are already produced by a host, the recursive sequence recombination techniques described above can be used to optimize production of the desired metabolic intermediate, including such features as increasing enzyme substrate specificity and turnover number, altering metabolic fluxes to reduce the concentrations of toxic substrates or intermediates, increasing resistance of the host to such toxic compounds, eliminating, reducing or altering the need for inducers of gene expression/activity, increasing the production of enzymes necessary for metabolism, etc.

Enzymes can also be evolved for improved activity in solvents other than water. This is useful because intermediates in chemical syntheses are often protected by blocking groups which dramatically affect the solubility of the compound in aqueous solvents. Many compounds can be produced by a combination of pure chemical and enzymically catalyzed reactions. Performing enzymic reactions on almost insoluble substrates is clearly very inefficient, so the availability of enzymes that are active in other solvents will be of great use. One example of such a scheme is the evolution of a para- nitrobenzyl esterase to remove protecting groups from an intermediate in loracarbef synthesis (Moore, J.C. and

Arnold, F.H. *Nature Biotechnology* 14:458-467 (1996)). In this case alternating rounds of error-prone PCR and colony screening for production of a fluorescent reporter from a substrate analogue were used to generate a mutant esterase that was 16-fold more active than the parent molecule in 30% dimethylformamide. No individual mutation was found to contribute more than a 2-fold increase in activity, but it was the combination of a number of mutations which led to the overall increase.

Structural analysis of the mutant protein showed that the amino acid changes were distributed throughout the length of the protein in a manner that could not have been rationally predicted. Sequential rounds of error-prone PCR have the problem that after each round all but one mutant is discarded, with a concomitant loss of information contained in all the other beneficial mutations. Recursive sequence recombination avoids this problem, and would thus be ideally suited to evolving enzymes for catalysis in other solvents, as well as in conditions where salt concentrations or pH were different from the original enzyme optima.

In addition, the yield of almost any metabolic pathway can be increased, whether consisting entirely of genes endogenous to the host organisms or all or partly heterologous genes. Optimization of the expression levels of the enzymes in a pathway is more complex than simply maximizing expression. In some cases regulation, rather than constitutive expression of an enzyme may be advantageous for cell growth and therefore for product yield, as seen for production of phenylalanine (Backman et al. *Ann. NY Acad. Sci.* 589:16-24 (1990)) and 2-keto-L- gluconic acid (Anderson et al. U.S. 5,032,514). In addition, it is often advantageous for industrial purposes to express proteins in organisms other than their original hosts. New host strains may be preferable for a variety of reasons, including ease of cloning and transformation, pathogenicity, ability to survive in particular environments and a knowledge of the physiology and genetics of the organisms. However, proteins expressed in heterologous organisms often show markedly reduced activity for a variety of reasons including inability to fold properly in the new host (Sarthy et al. *Appl. Environ. Micro.* 53:1996-2000 (1987)). Such difficulties can indeed be overcome by the recursive sequence recombination strategies of the instant invention.



## 8.14.2 ANTIBIOTICS

The range of natural small molecule antibiotics includes but is not limited to peptides, peptidolactones, thiopeptides, beta-lactams, glycopeptides, lantibiotics, microcins, polyketide-derived antibiotics (anthracyclins, tetracyclins, macrolides, avermectins, polyethers and ansamycins), chloramphenicol, aminoglycosides, aminocyclitols, polyoxins, agrocins and isoprenoids. There are at least three ways in which recursive sequence recombination techniques of the instant invention can be used to facilitate novel drug synthesis, or to improve biosynthesis of existing antibiotics.

First, antibiotic synthesis enzymes can be "evolved" together with transport systems that allow entry of compounds used as antibiotic precursors to improve uptake and incorporation of function-altering artificial side chain precursors. For example, penicillin V is produced by feeding *Penicillium* the artificial side chain precursor phenoxyacetic acid, and LY146032 by feeding *Streptomyces roseosporus* decanoic acid (Hopwood, Phil. Trans. R. Soc. Lond. B 324:549-562 (1989)). Poor precursor uptake and poor incorporation by the synthesizing enzyme often lead to inefficient formation of the desired product. Recursive sequence recombination of these two systems can increase the yield of desired product.

Furthermore, a combinatorial approach can be taken in which an enzyme is stochastic &/or non-stochastic mutagenized for novel catalytic activity/substrate recognition (perhaps by including randomizing oligonucleotides in key positions such as the active site). A number of different substrates (for example, analogues of side chains that are normally incorporated into the antibiotic) can then be tested in combination with all the different enzymes and tested for biological activity. In this embodiment, plates are made containing different potential antibiotic precursors (such as the side chain analogues). The microorganisms containing the stochastic &/or non-stochastic mutagenized library (the library strain) are replicated onto those plates, together with a competing, antibiotic sensitive, microorganism (the indicator strain). Library cells that are able to incorporate the new side chain to produce an effective antibiotic will thus be able to compete with the indicator strain, and will be selected for.

Second, the expression of heterologous genes transferred from one antibiotic synthesizing organism to another can be optimized. The newly introduced enzyme(s) act on secondary metabolites in the host cell, transforming them into new compounds with novel properties. Using traditional methods, introduction of foreign genes into antibiotic synthesizing hosts has already resulted in the production of novel hybrid antibiotics. Examples include meder rhodin, dihydrogranatirhodin, 6-deoxyerythromycin A, isovalerylspiramycin and other hybrid macrolides (Cameron et. al. Appl. Biochem. Biotechnol. 38:105-140 (1993)). The recursive sequence recombination techniques of the instant invention can be used to optimize expression of the foreign genes, to stabilize the enzyme in the new host cell, and to increase the activity of the introduced enzyme against its new substrates in the new host cell. In some embodiments of the invention, the host genome may also be so optimized.

Third, the substrate specificity of an enzyme involved in secondary metabolism can be altered so that it will act on and modify a new compound or so that its activity is changed and it acts at a different subset of positions of its normal substrate. Recursive sequence recombination can be used to alter the substrate specificities of enzymes. Furthermore, in addition to recursive sequence recombination of individual enzymes being a strategy to generate novel antibiotics, recursive sequence recombination of entire pathways, by altering enzyme ratios, will alter metabolite fluxes and may result, not only in increased antibiotic synthesis, but also in the synthesis of different antibiotics. This can be deduced from the observation that expression of different genes from the same cluster in a foreign host leads to different products being formed (see p. 80 in Hutchinson et. al., (1991) Ann NY Acad Sci, 646:78-93).

Recursive sequence recombination of the introduced gene clusters may result in a variety of expression levels of different proteins within the cluster (because it produces different combinations of, in this case regulatory, mutations). This in turn may lead to a variety of different end products. Thus, "evolution" of an existing antibiotic synthesizing pathway could be used to generate novel antibiotics either by modifying the rates or substrate specificities of enzymes in that pathway.

Additionally, antibiotics can also be produced in vitro by the action of a purified enzyme on a precursor. For example isopenicillin N synthase catalyses the cyclization of many analogues of its normal substrate (d-(L- $\alpha$ -aminoadipyl)-L-cysteinyl-D-valine) (Hutchinson, Med. Res. Rev. 8:557-567 (1988)). Many of these products are active as antibiotics. A wide variety of substrate analogues can be tested for incorporation by secondary metabolite synthesizing enzymes without concern for the initial efficiency of the reaction. Recursive sequence recombination can be used subsequently to increase the rate of reaction with a promising new substrate.

Thus, organisms already producing a desired antibiotic can be evolved with the recursive sequence recombination techniques described above to maximize production of that antibiotic. Additionally, new antibiotics can be evolved by manipulation of genetic material from the host by the recursive sequence recombination techniques described above. Genes for antibiotic production can be transferred to a preferred host after cycles of recursive sequence recombination or can be evolved in the preferred host as described above.

Antibiotic genes are generally clustered and are often positively regulated, making them especially attractive candidates for the recursive sequence recombination techniques of the instant invention. Additionally, some genes of related pathways show cross-hybridization, making them preferred candidates for the generation of new pathways for new antibiotics by the recursive sequence recombination techniques of the invention.

Furthermore, increases in secondary metabolite production including enhancement of substrate fluxes (by increasing the rate of a rate limiting enzyme, deregulation of the pathway by suppression of negative control elements or over expression of activators and the relief of feedback controls by mutation of the regulated enzyme to a feedback-insensitive deregulated protein) can be achieved by recursive sequence recombination without exhaustive analysis of the regulatory mechanisms governing expression of the relevant gene clusters.

The host chosen for expression of evolved genes is preferably resistant to the antibiotic produced, although in some instances production methods can be designed so as to sacrifice host cells when the amount of antibiotic produced is commercially significant yet lethal to the host. Similarly, bioreactors can be designed so that the growth medium is continually replenished, thereby "drawing off" antibiotic produced and sparing the lives of the producing cells. Preferably, the mechanism of resistance is not the degradation of the antibiotic produced.

Numerous screening methods for increased antibiotic expression are known in the art, as discussed above, including screening for organisms that are more resistant to the antibiotic that they produce. This may result from linkage between expression of the antibiotic synthesis and antibiotic resistance genes (Chater, *Bio/Technology* 8:115-121 (1990)). Another screening method is to fuse a reporter gene (e.g. *xylE* from the *Pseudomonas* TOL plasmid) to the antibiotic production genes. Antibiotic synthesis gene expression can then be measured by looking for expression of the reporter (e.g. *xylE* encodes a catechol dioxygenase which produces yellow muconic semialdehyde when colonies are sprayed with catechol (Zukowski et al. *Proc. Natl. Acad. Sci. U.S.A.* 80:1101-1105 (1983)). The wide variety of cloned antibiotic genes provides a wealth of starting materials for the recursive sequence recombination techniques of the instant invention. For example, genes have been cloned from *Streptomyces cattleya* which direct cephamycin C synthesis in the non-antibiotic producer *Streptomyces lividans* (Chen et al. *Bio/Technology* 6:1222-1224 (1988)). Clustered genes for penicillin biosynthesis (-(L- $\alpha$ -aminoadipyl)-L-cysteinyl-D-valine synthetase; isopenicillin N synthetase and acyl coenzyme A:6-aminopenicillanic acid acyltransferase) have been cloned from *Penicillium chrysogenum*. Transfer of these genes into *Neurospora crassa* and *Aspergillus niger* result in the synthesis of active penicillin V (Smith et al. *Bio/Technology* 8:39-41 (1990)). For a review of cloned genes involved in Cephalosporin C, Penicillins G and V and Cephamycin C biosynthesis, see Piepersberg, *Crit. Rev. Biotechnol.* 14:251-285 (1994). For a review of cloned clusters of antibiotic-producing genes, see Chater *Bio/Technology* 8:115-121 (1990). Other examples of antibiotic synthesis genes transferred to industrial producing strains, or over expression of genes, include tylosin, cephamycin C, cephalosporin C, LL-E33288 complex (an antitumor and antibacterial agent), doxorubicin, spiramycin and other

macrolide antibiotics, reviewed in Cameron et al. *Appl. Biochem. Biotechnol.* 38:105-140 (1993).

#### 8.14.3 BIOSYNTHESIS TO REPLACE CHEMICAL SYNTHESIS OF ANTIBIOTICS

Some antibiotics are currently made by chemical modifications of biologically produced starting compounds. Complete biosynthesis of the desired molecules may currently be impractical because of the lack of an enzyme with the required enzymatic activity and substrate specificity. For example, 7-aminodeacetoxycephalosporanic acid (7-ADCA) is a precursor for semi-synthetically produced cephalosporins. 7-ADCA is made by a chemical ring expansion from penicillin V followed by enzymatic deacylation of the phenoxyacetal group. Cephalosporins could in principle be produced biologically from their corresponding penicillins (e.g., cephalosporin V or G from penicillin V or G) using penicillin N expandase, but other penicillins (such as penicillin V or G) are not used as substrates by known expandases. The recursive sequence recombination techniques of the invention can be used to alter the enzyme so that it will use penicillin V as a substrate. Similarly, penicillin transacylase could be so modified to accept cephalosporins or cephamycins as substrates.

In yet another example, penicillin amidase expressed in *E. coli* is a key enzyme in the production of penicillin G derivatives. The enzyme is generated from a precursor peptide and tends to accumulate as insoluble aggregates in the periplasm unless non-metabolizable sugars are present in the medium (Scherrer et al. , *Appl. Microbiol. Biotechnol.* 42:85-91 (1994)). Evolution of this enzyme through the methods of the instant invention could be used to generate an enzyme that folds better, leading to a higher level of active enzyme expression.

In yet another example, Penicillin G acylase covalently linked to agarose is used in the synthesis of penicillin G derivatives. The enzyme can be stabilized for increased activity, longevity and/or thermal stability by chemical modification (Fernandez-Lafuente et. al. *Enzyme Microb. Technol.* 14:489-495 (1992)). Increased thermal stability is an

especially attractive application of the recursive sequence recombination techniques of the instant invention, which can obviate the need for the chemical modification of such enzymes. Selection for thermostability can be performed *in vivo* in *E. coli* or in thermophiles at higher temperatures. In general, thermostability is a good first step in enhancing general stabilization of enzymes. Random mutagenesis and selection can also be used to adapt enzymes to function in non-aqueous solvents (Arnold Curr Opin Biotechnol, 4:450-455 (1993); Chen et. al. Proc. Natl. Acad. Sci. U.S.A., 90:5618-5622 (1993)). Recursive sequence recombination represents a more powerful (since recombinogenic) method of generating mutant enzymes that are stable and active in non-aqueous environments. Additional screening can be done on the basis of enzyme stability in solvents.

#### 8.14.4 POLYKETIDES

Polyketides include antibiotics such as tetracycline and erythromycin, anti-cancer agents such as daunomycin, immunosuppressants such as FK506 and rapamycin and veterinary products such as monesin and avermectin. Polyketide synthases (PKS's) are multifunctional enzymes that control the chain length, choice of chain-building units and reductive cycle that generates the huge variation in naturally occurring polyketides. Polyketides are built up by sequential transfers of "extender units" (fatty acyl CoA groups) onto the appropriate starter unit (examples are acetate, coumarate, propionate and malonamide). The PKS's determine the number of condensation reactions and the type of extender groups added and may also fold and cyclize the polyketide precursor. PKS's reduce specific  $\beta$ -keto groups and may dehydrate the resultant  $\beta$ -hydroxyls to form double bonds. Modifications of the nature or number of building blocks used, positions at which  $\beta$ -keto groups are reduced, the extent of reduction and different positions of possible cyclizations, result in formation of different final products. Polyketide research is currently focused on modification and inhibitor studies, site directed mutagenesis and 3-D structure elucidation to lay the groundwork for rational changes in enzymes that will lead to new polyketide products.

Recently, McDaniel et al. (Science 262:1546- 1550 (1995)) have developed a *Streptomyces* host-vector system for efficient construction and expression of recombinant PKSs. Hutchinson (Bio/Technolo 12:375-308 (1994)) reviewed targeted mutation of specific biosynthetic genes and suggested that microbial isolates can be screened by DNA hybridization for genes associated with known pharmacologically active agents so as to provide new metabolites or increased yields of metabolites already being produced. In particular, that review focuses on polyketide synthase and pathways to aminoglycoside and oligopeptide antibiotics.

The recursive sequence recombination techniques of the instant invention can be used to generate modified enzymes that produce novel polyketides without such detailed analytical effort. The availability of the PKS genes on plasmids and the existence of *E. coli*- *Streptomyces* shuttle vectors (Wehmeier Gene 165:149-150 (1995)) makes the process of recursive sequence recombination especially attractive by the techniques described above. Techniques for selection of antibiotic producing organisms can be used as described above; additionally, in some embodiments screening for a particular desired polyketide activity or compound is preferable.

#### 8.14.5 ISOPRENOIDS

Isoprenoids result from cyclization of farnesyl pyrophosphate by sesquiterpene synthases. The diversity of isoprenoids is generated not by the backbone, but by control of cyclization. Cloned examples of isoprenoid synthesis genes include trichodiene synthase from *Fusarium sporotrichioides*, pentalene synthase from *Streptomyces*, aristolochene synthase from *Penicillium roquefortii*, and epi-aristolochene synthase from *N. tabacum* (Cane, D.E. (1995). Isoprenoid antibiotics, pages 633-655, in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L.C. & Stuttard, C., published by Butterworth-Heinemann). Recursive sequence recombination of sesquiterpene synthases will be of use both in allowing expression of these enzymes in heterologous hosts (such as plants and industrial microbial strains) and in alteration of enzymes to change the cyclized product

made. A large number of isoprenoids are active as antiviral, antibacterial, antifungal, herbicidal, insecticidal or cytostatic agents. Antibacterial and antifungal isoprenoids could thus be preferably screened for using the indicator cell type system described above, with the producing cell competing with bacteria or fungi for nutrients. Antiviral isoprenoids could be screened for preferably by their ability to confer resistance to viral attack on the producing cell.

#### 8.14.6 BIOACTIVE PEPTIDE DERIVATIVES

Examples of bioactive non-ribosomally synthesized peptides include the antibiotics cyclosporin, pepstatin, actinomycin, gramicidin, depsipeptides, vancomycin, etc. These peptide derivatives are synthesized by complex enzymes rather than ribosomes. Again, increasing the yield of such non-ribosomally synthesized peptide antibiotics has thus far been done by genetic identification of biosynthetic "bottlenecks" and over expression of specific enzymes (See, for example, p. 133-135 in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L.C. & Stuttard, C., published by Butterworth-Heinemann). Recursive sequence recombination of the enzyme clusters can be used to improve the yields of existing bioactive non-ribosomally made peptides in both natural and heterologous hosts.

Like polyketide synthases, peptide synthases are modular and multifunctional enzymes catalyzing condensation reactions between activated building blocks (in this case amino acids) followed by modifications of those building blocks (see Kleinkauf, H. and von Dohren, H. Eur. J. Biochem. 236:335-351 (1996)). Thus, as for polyketide synthases, recursive sequence recombination can also be used to alter peptide synthases: modifying the specificity of the amino acid recognized by each binding site on the enzyme and altering the activity or substrate specificities of sites that modify these amino acids to produce novel compounds with antibiotic activity. Other peptide antibiotics are made ribosomally and then post-translationally modified. Examples of this type of antibiotics are lantibiotics (produced by gram positive bacteria such as *Staphylococcus*, *Streptomyces*, *Bacillus*, and *Actinoplanes*) and microcins (produced by *Enterobacteriaceae*).



Modifications of the original peptide include (in lantibiotics) dehydration of serine and threonine, condensation of dehydroamino acids with cysteine, or simple N- and C-terminal blocking (microcins). For ribosomally made antibiotics both the peptide-encoding sequence and the modifying enzymes may have their expression levels modified by recursive sequence recombination. Again, this will lead to both increased levels of antibiotic synthesis, and by modulation of the levels of the modifying enzymes (and the sequence of the ribosomally synthesized peptide itself) novel antibiotics.

Screening can be done as for other antibiotics as described above, including competition with a sensitive (or even initially insensitive) microbial species. Use of competing bacteria that have resistances to the antibiotic being produced will select strongly either for greatly elevated levels of that antibiotic (so that it swamps out the resistance mechanism) or for novel derivatives of that antibiotic that are not neutralized by the resistance mechanism.

#### 8.14.7 POLYMERS

Several examples of metabolic engineering to produce biopolymers have been reported, including the production of the biodegradable plastic polyhydroxybutyrate (PHB), and the polysaccharide xanthan gum. For a review, see Cameron et al. *Applied Biochem. Biotech.* 38:105-140 (1993). Genes for these pathways have been cloned, making them excellent candidates for the recursive sequence recombination techniques described above. Expression of such evolved genes in a commercially viable host such as *E. coli* is an especially attractive application of this technology.

Examples of starting materials for recursive sequence recombination include but are not limited to genes from bacteria such as *Alcaligenes*, *Zoogloea*, *Rhizobium*, *Bacillus*, and *Azobacter*, which produce polyhydroxyalkanoates (PHAs) such as polyhydroxybutyrate (PHB) intracellularly as energy reserve materials in response to stress. Genes from *Alcaligenes eutrophus* that encode enzymes catalyzing the conversion of acetoacetyl CoA to PHB have been transferred both to *E. coli* and to the plant *Arabidopsis thaliana* (Poirier et al. *Science* 256:520-523 (1992)). Two of these genes (*phbB* and *phbC*, encoding

acetoacetyl-CoA reductase and PHB synthase respectively) allow production of PHE in *Arabidopsis*. The plants producing the plastic are stunted, probably because of adverse interactions between the new metabolic pathway and the plants' original metabolism (i.e., depletion of substrate from the mevalonate pathway). Improved production of PHB in plants has been attempted by localization of the pathway enzymes to organelles such as plastids. Other strategies such as regulation of tissue specificity, expression timing and cellular localization have been suggested to solve the deleterious effects of PHB expression in plants. The recursive sequence recombination techniques of the invention can be used to modify such heterologous genes as well as specific cloned interacting pathways (e.g., mevalonate), and to optimize PHB synthesis in industrial microbial strains, for example to remove the requirement for stresses (such as nitrogen limitation) in growth conditions.

Additionally, other microbial polyesters are made by different bacteria in which additional monomers are incorporated into the polymer (Peoples et al. in *Novel Biodegradable Microbial Polymers*, EA Dawes, ed., pp191-202 (1990) ). Recursive sequence recombination of these genes or pathways singly or in combination into a heterologous host will allow the production of a variety of polymers with differing properties, including variation of the monomer subunit ratios in the polymer.

Another polymer whose synthesis may be manipulated by recursive sequence recombination is cellulose. The genes for cellulose biosynthesis have been cloned from *Agrobacterium tumefaciens* (Matthysse, A.G. et. al. *J. Bacteriol.* 177:1069-1075 (1995)). Recursive sequence recombination of this biosynthetic pathway could be used either to increase synthesis of cellulose, or to produce mutants in which alternative sugars are incorporated into the polymer.

#### 8.14.8 CAROTENOIDS

Carotenoids are a family of over 600 terpenoids produced in the general isoprenoid biosynthetic pathway by bacteria, fungi and plants (for a review, see Armstrong, *J. Bact.* 176:4795-4802 (1994)). These pigments protect organisms against photooxidative damage

as well as functioning as anti-tumor agents, free radical-scavenging anti-oxidants, and enhancers of the immune response. Additionally, they are used commercially in pigmentation of cultured fish and shellfish. Examples of carotenoids include but are not limited to myxobacton, spheroidene, spheroidenone, lutein, astaxanthin, violaxanthin, 4-ketorulene, myxoxanthrophyll, echinenone, lycopene, zeaxanthin and its mono- and di-glucosides, alpha-, beta-, gamma- and sigma-carotene, beta-cryptoxanthin monoglucoside and neoxanthin.

Carotenoid synthesis is catalyzed by relatively small numbers of clustered genes: 11 different genes within 12 kb of DNA from *Myxococcus xanthus* (Botella et al. Eur. J. Biochem. 233:238-248 (1995)) and 8 genes within 9 kb of DNA from *Rhodobacter sphaeroides* (Lang et. al. J. Bact. 177:2064-2073 (1995)). In some microorganisms, such as *Thermus thermophilus*, these genes are plasmid-borne (Tabata et al. FEBS Letts 341:251-255 (1994)). These features make carotenoid synthetic pathways especially attractive candidates for recursive sequence recombination.

Transfer of some carotenoid genes into heterologous organisms results in expression. For example, genes from *Erwinia uredovora* and *Haematococcus pluvialis* will function together in *E. coli* (Kajiware et al. Plant Mol. Biol. 29:343-352 (1995)). *E. herbicola* genes will function in *R. sphaeroides* (Hunter et al. J. Bact. 176:3692-3697 (1994)). However, some other genes do not; for example, *R. capsulatus* genes do not direct carotenoid synthesis in *E. coli* (Marrs, J. Bact. 146:1003-1012 (1981)).

In an embodiment of the invention, the recursive sequence recombination techniques of the invention can be used to generate variants in the regulatory and/or structural elements of genes in the carotenoid synthesis pathway, allowing increased expression in heterologous hosts. Indeed, traditional techniques have been used to increase carotenoid production by increasing expression of a rate limiting enzyme in *Thermus thermophilus* (Hoshino et al. Appl. Environ. Micro. 59:3150-3153 (1993)). Furthermore, mutation of regulatory genes can cause constitutive expression of carotenoid synthesis in actinomycetes, where carotenoid photoinducibility is otherwise unstable and lost at a relatively high frequency in some species (Kato et al. Mol. Gen. Genet. 247:387-390 (1995)). These are both mutations that can be obtained by recursive sequence

recombination.

The recursive sequence recombination techniques of the invention as described above can be used to evolve one or more carotenoid synthesis genes in a desired host without the need for analysis of regulatory mechanisms. Since carotenoids are colored, a colorimetric assay in microtiter plates, or even on growth media plates, can be used for screening for increased production.

In addition to increasing expression of carotenoids, carotenogenic biosynthetic pathways have the potential to produce a wide diversity of carotenoids, as the enzymes involved appear to be specific for the type of reaction they will catalyze, but not for the substrate that they modify. For example, two enzymes from the marine bacterium *Agrobacterium aurantiacum* (CrtW and CrtZ) synthesize six different ketocarotenoids from beta-carotene (Misawa et al. J. Bact. 177:6576-6584 (1995)). This relaxed substrate specificity means that a diversity of substrates can be transformed into an even greater diversity of products. Introduction of foreign carotenoid genes into a cell can lead to novel and functional carotenoid-protein complexes, for example in photosynthetic complexes (Hunter et al. J.Bact. 176:3692- 3697 (1994)). Thus, the deliberate recombination of enzymes through the recursive sequence recombination techniques of the invention is likely to generate novel compounds. Screening for such compounds can be accomplished, for example, by the cell competition/survival techniques discussed above and by a colorimetric assay for pigmented compounds.

Another method of identifying new compounds is to use standard analytical techniques such as mass spectroscopy, nuclear magnetic resonance, high performance liquid chromatography, etc. Recombinant microorganisms can be pooled and extracts or media supernatants assayed from these pools. Any positive pool can then be subdivided and the procedure repeated until the single positive is identified ("sib-selection").

### 8.14.9 INDIGO BIOSYNTHESIS

Many dyes, i.e. agents for imparting color, are specialty chemicals with significant markets. As an example, indigo is currently produced chemically. However, nine genes have been combined in *E. coli* to allow the synthesis of indigo from glucose via the tryptophan/indole pathway (Murdock et al. *Bio/Technology* 11:381-386 (1993)). A number of manipulations were performed to optimize indigo synthesis: cloning of nine genes, modification of the fermentation medium and directed changes in two operons to increase reaction rates and catalytic activities of several enzymes.

Nevertheless, bacterially produced indigo is not currently an economic proposition. The recursive sequence recombination techniques of the instant invention could be used to optimize indigo synthesizing enzyme expression levels and catalytic activities, leading to increased indigo production, thereby making the process commercially viable and reducing the environmental impact of indigo manufacture. Screening for increased indigo production can be done by colorimetric assays of cultures in microtiter plates.

### 8.14.10 AMINO ACIDS

Amino acids of particular commercial importance include but are not limited to phenylalanine, monosodium glutamate, glycine, lysine, threonine, tryptophan and methionine. Backman et al. (*Ann. NY Acad. Sci.* 589:16-24 (1990)) disclosed the enhanced production of phenylalanine in *E. coli* via a systematic and downstream strategy covering organism selection, optimization of biosynthetic capacity, and development of fermentation and recovery processes. As described in Simpson et al. (*Biochem Soc Trans.* 23:381-387 (1995)), current work in the field of amino acid production is focused on understanding the regulation of these pathways in great molecular detail.

The recursive sequence recombination techniques of the instant invention would obviate the need for this analysis to obtain bacterial strains with higher secreted amino acid yields. Amino acid production could be optimized for expression using recursive sequence recombination of the amino acid synthesis and secretion genes as well as

enzymes at the regulatory phosphoenolpyruvate branchpoint, from such organisms as *Serratia marcescens*, *Bacillus*, and the *Corynebacterium* -*Brevibacterium* group. In some embodiments of the invention, screening for enhanced production is preferably done in microtiter wells, using chemical tests well known in the art that are specific for the desired amino acid. Screening/selection for amino acid synthesis can also be done by using auxotrophic reporter cells that are themselves unable to synthesize the amino acid in question. If these reporter cells also produce a compound that stimulates the growth of the amino acid producer (this could be a growth factor, or even a different amino acid), then library cells that produce more amino acid will in turn receive more growth stimulant and will therefore grow more rapidly.

#### 8.14.11 VITAMIN C SYNTHESIS

L-Ascorbic acid (vitamin C) is a commercially important vitamin with a world production of over 35,000 tons in 1984. Most vitamin C is currently manufactured chemically by the Reichstein process, although recently bacteria have been engineered that are able to transform glucose to 2,5-keto-gluconic acid, and that product to 2-keto-L-idonic acid, the precursor to L-ascorbic acid (Boudrant, *Enzyme Microb. Technol.* 12:322-329 (1990)).

The efficiencies of these enzymatic steps in bacteria are currently low. Using the recursive sequence recombination techniques of the instant invention, the genes can be genetically engineered to create one or more operons followed by expression optimization of such a hybrid L-ascorbic acid synthetic pathway to result in commercially viable microbial vitamin C biosynthesis. In some embodiments, screening for enhanced L-ascorbic acid production is preferably done in microtiter plates, using assays well known in the art.

## 8.15 TEST FOR RESISTANCE TO DRUGS

### 8.15.1 FIND DRUGS THAT INDUCE RESISTANCE SLOWLY

A similar strategy can be used to simulate viral acquisition of drug resistance. The object is to identify drugs for which resistance can be acquired only slowly, if at all. The viruses to be evolved are those that cause infections in humans for which at least modestly effective drugs are available. Substrates for recombination can come from induced mutants, natural variants of the same viral strain or different viruses. If the target of the drug is known (e.g., nucleotide analogs which inhibit the reverse transcriptase gene of HIV), focused libraries containing variants of the target gene can be produced. Recombination of a viral genome with a library of fragments is usually performed *in vitro*. However, in situations in which the library of fragments constitutes variants of viral genomes or fragments that can be encompassed in such genomes, recombination can also be performed *in vivo*, e.g., by transfecting cells with multiple substrate copies (see Section V). For screening, recombinant viral genomes are introduced into host cells susceptible to infection by the virus and the cells are exposed to a drug effective against the virus (initially at low concentration). The cells can be spun to remove any noninfected virus. After a period of infection, progeny viruses can be collected from the culture medium, the progeny viruses being enriched for viruses that have acquired at least partial resistance to the drug. Alternatively, virally infected cells can be plated in a soft agar lawn and resistant viruses isolated from plaques. Plaque size provides some indication of the degree of viral resistance.

Progeny viruses surviving screening are subject to additional rounds of recombination and screening at increased stringency until a predetermined level of drug resistance has been acquired. The predetermined level of drug resistance may reflect the maximum dosage of a drug practical to administer to a patient without intolerable side effects. The analysis is particularly valuable for investigating acquisition of resistance to various combination of drugs, such as the growing list of approved anti-HIV drugs (e.g., AZT, ddI, ddC, d4T, TIBO 82150, nevirapine, 3TC, crixivan and ritonavir).

### 8.15.2 METHOD TO EVOLVE YEAST STRAINS

Fragments are cloned into a YAC vector, and the resulting YAC library is transformed into competent yeast cells. Transformants containing a YAC are identified by selecting for a positive selection marker present on the YAC. The cells are allowed to recover and are then pooled. Thereafter, the cells are induced to sporulate by transferring the cells from rich medium, to nitrogen and carbon limiting medium. In the course of sporulation, cells undergo meiosis. Spores are then induced to mate by return to rich media. Optionally, asci are lysed to liberate spores, so that the spores can mate with other spores originating from other asci. Mating results in recombination between YACs bearing different inserts, and between YACs and natural yeast chromosomes. The latter can be promoted by irradiating spores with ultra violet light. Recombination can give rise to new phenotypes either as a result of genes expressed by fragments on the YACs or as a result of recombination with host genes, or both.

After induction of recombination between YACs and natural yeast chromosomes, YACs are often eliminated by selecting against a negative selection marker on the YACs. For example, YACs containing the marker URA3 can be selected against by propagation on media containing 5-fluoro orotic acid. Any exogenous or altered genetic material that remains is contained within natural yeast chromosomes. Optionally, further rounds of recombination between natural yeast chromosomes can be performed after elimination of YACs. Optionally, the same or different library of YACs can be transformed into the cells, and the above steps repeated. By recursively repeating this process, the diversity of the population is increased prior to screening.

After elimination of YACs, yeast are then screened or selected for a desired property. The property can be a new property conferred by transferred fragments, such as production of an antibiotic. The property can also be an improved property of the yeast such as improved capacity to express or secrete an exogenous protein, improved recombinogenicity, improved stability to temperature or solvents, or other property required of commercial or research strains of yeast.



Yeast strains surviving selection/screening are then subject to a further round of recombination. Recombination can be exclusively between the chromosomes of yeast surviving selection/screening. Alternatively, a library of fragments can be introduced into the yeast cells and recombined with endogenous yeast chromosomes as before. This library of fragments can be the same or different from the library used in the previous round of transformation. For example, the YACs could contain a library of genomic DNA isolated from a pool of the improved strains obtained in the earlier steps. YACs are eliminated as before, followed by additional rounds of recombination and/or transformation with further YAC libraries. Recombination is followed by another round of selection/screening, as above.

Further rounds of recombination/screening can be performed as needed until a yeast strain has evolved to acquire the desired property.

An exemplary scheme for evolving yeast by introduction of a YAC library is yeast containing an endogenous diploid genome and a YAC library of fragments representing variants of a sequence. The library is transformed into the cells to yield 100-1000 colonies per  $\mu\text{g}$  DNA. Most transformed yeast cells now harbor a single YAC as well as endogenous chromosomes. Meiosis is induced by growth on nitrogen and carbon limiting medium. In the course of meiosis the YACs recombine with other chromosomes in the same cell. Haploid spores resulting from meiosis mate and regenerated diploid forms. The diploid forms now harbor recombinant chromosomes, parts of which come from endogenous chromosomes and parts from YACs.

Optionally, the YACs can now be cured from the cells by selecting against a negative selection marker present on the YACS. Irrespective whether YACS are selected against, cells are then screened or selected for a desired property. Cells surviving selection/screening are transformed with another YAC library to start another stochastic &/or non-stochastic mutagenesis cycle.

### 8.15.3 EVOLVE YACs FOR TRANSFER INTO RECIPIENT STRAIN

These methods are based in part on the fact that multiple YACs can be harbored in the same yeast cell, and YAC-YAC recombination is known to occur (Green & Olson, Science 250, 94-98 1990)). Inter-YAC recombination provides a format for which families of homologous genes harbored on fragments of >20 kb can be stochastic &/or non-stochastic mutagenized in vivo.

The starting population of DNA fragments show sequence similarity with each other but differ as a result of for example, induced, allelic or species diversity. Often DNA fragments are known or suspected to encode multiple genes that function in a common pathway.

The fragments are cloned into a Yac and transformed into yeast, typically with positive selection for transformants. The transformants are induced to sporulate, as a result of which chromosomes undergo meiosis. The cells are then mated. Most of the resulting diploid cells now carry two YACs each having a different insert. These are again induced to sporulate and mated. The resulting cells harbor YACs of recombined sequence. The cells can then be screened or selected for a desired property. Typically, such selection occurs in the yeast strain used for stochastic &/or non-stochastic mutagenesis. However, if fragments being stochastic &/or non-stochastic mutagenized are not expressed in yeast, YACs can be isolated and transferred to an appropriate cell type in which they are expressed for screening. Examples of such properties include the synthesis or degradation of a desired compound, increased secretion of a desired gene product, or other detectable phenotype.

Preferably, the YAC library is transformed into haploid a and haploid a cells. These cells are then induced to mate with each other, i.e., they are pooled and induced to mate by growth on rich medium. The diploid cells, each carrying two YACs, are then transferred to sporulation medium. During sporulation, the cells undergo meiosis, and homologous chromosomes recombine. In this case, the genes harbored in the YACs will recombine, diversifying their sequences. The resulting haploid ascospores are then liberated

from the asci by enzymatic degradation of the asci wall or other available means and the pooled liberated haploid aco-spores are induced to mate by transfer to rich medium. This process is repeated for several cycles to increase the diversity of the DNA cloned into the YACs. The resulting population of yeast cells, preferably in the haploid state, are either screened for improved properties, or the diversified DNA is delivered to another host cell or organism for screening.

Cells surviving selection/screening are subjected to successive cycles of pooling, sporulation, mating and selection/screening until the desired phenotype has been observed. Recombination can be achieved simply by transferring cells from rich medium to carbon and nitrogen limited medium to induce sporulation, and then returning the spores to rich media to induce mating. Asci can be lysed to stimulate mating of spores originating from different asci.

After YACs have been evolved to encode a desired property they can be transferred to other cell types. Transfer can be by protoplast fusion, or retransformation with isolated DNA. For example, transfer of YACs from yeast to mammalian cells is discussed by Monaco & Larin, *Trends in Biotechnology* 12, 280-286 (1994); Montoliu et al., *Reprod. Fertil. Dev.* 6, 577-84 (1994); Lamb et al., *Curr. Opin. Genet. Dev.* 5, 342-8 (1995). An exemplary scheme for stochastic &/or non-stochastic mutagenesis a YAC fragment library in yeast is shown herein. A library of YAC fragments representing genetic variants are transformed into yeast that have diploid endogenous chromosomes. The transformed yeast continue to have diploid endogenous chromosomes, plus a single YAC. The yeast are induced to undergo meiosis and sporulate. The spores contain haploid genomes and are selected for those which contain a YAC, using the YAC selective marker. The spores are induced to mate generating diploid cells. The diploid cells now contain two YACs bearing different inserts as well as diploid endogenous chromosomes. The cells are again induced to undergo meiosis and sporulate. during meiosis, recombination occurs between the YAC inserts, and recombinant YACs are segregated to ascocytes. Some ascocytes thus contain haploid endogenous chromosomes plus a YAC chromosome with a recombinant insert. The ascocytes mature to spores, which can mate

again generating diploid cells. Some diploid cells now possess a diploid complement of endogenous chromosomes plus two recombinant YACs. These cells can then be taken through further cycles of meiosis, sporulation and mating. In each cycle, further recombination occurs between YAC inserts and further recombinant forms of inserts are generated. After one or several cycles of recombination has occurred, cells can be tested for acquisition of a desired property. Further cycles of recombination, followed by selection, can then be performed in similar fashion.

#### **8.15.4 IN VIVO STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF GENES BY THE RECURSIVE MATING OF YEAST CELLS HARBORING HOMOLOGOUS GENES IN IDENTICAL LOCI**

A goal of DNA stochastic &/or non-stochastic mutagenesis is to mimic and expand the combinatorial capabilities of sexual recombination. In vitro DNA stochastic &/or non-stochastic mutagenesis succeeds in this process. However, by changing the mechanism of recombination and altering the conditions under which recombination occurs, naturally in vitro recombination methods may jeopardize intrinsic information in a DNA sequence that renders it "evolvable." Stochastic &/or non-stochastic mutagenesis in vivo by employing the natural crossing over mechanisms that occur during meiosis may access inherent natural sequence information and provide a means of creating higher quality stochastic &/or non-stochastic mutagenized libraries. Described here is a method for the in vivo stochastic &/or non-stochastic mutagenesis of DNA that utilizes the natural mechanisms of meiotic recombination and provides an alternative method for DNA stochastic &/or non-stochastic mutagenesis.

The basic strategy is to clone genes to be stochastic &/or non-stochastic mutagenized into identical loci within the haploid genome of yeast. The haploid cells are then recursively induced to mate and to sporulate. The process subjects the cloned genes to recursive recombination during recursive cycles of meiosis. The resulting stochastic &/or non-stochastic mutagenized genes are then screened in situ or isolated and screened under different conditions.

For example, if one wished to reassemble a family of five lipase genes, the following provides a means of doing so in vivo.

The open reading frame of each lipase is amplified by the PCR such that each ORF is flanked by identical 3' and 5' sequences. The 5' flanking sequence is identical to a region within the 5' coding sequence of the *S. cerevisiae* *ura 3* gene and the 3' flanking sequence is identical to a region within the 3' of the *ura 3* gene. The flanking sequences are chosen such that homologous recombination of the PCR product with the *ura 3* gene results in the incorporation of the lipase gene and the disruption of the *ura 3* ORF. Both *S. cerevisiae* *a* and *h* haploid cells are then transformed with each of the PCR amplified lipase ORFs, and cells having incorporated a lipase gene into the *ura 3* locus are selected by growth on 5 fluoro orotic acid (5FOA is lethal to cells expressing functional *URA3*). The result is 10 cell types, two different mating types each harboring one of the five lipase genes in the disrupted *ura 3* locus. These cells are then pooled and grown under conditions where mating between the *a* and *h* cells are favored, e.g. in rich medium. Mating results in a combinatorial mixture of diploid cells having all 32 possible combinations of lipase genes in the two *ura 3* loci. The cells are then induced to sporulate by growth under carbon and nitrogen limited conditions. During sporulation the diploid cells undergo meiosis to form four (two *a* and two *h*) haploid ascospores housed in an ascus.

During meiosis II of the sporulation process sister chromatids align and crossover. The lipase genes cloned into the *ura 3* loci will also align and recombine. Thus the resulting haploid ascospores will represent a library of cells each harboring a different possible chimeric lipase gene, each a unique result of the meiotic recombination of the two lipase genes in the original diploid cell. The walls of asci are degraded by treatment with zymolase to liberate and allow the mixing of the individual ascospores. This mixture is then grown under conditions that promote the mating of the *a* and *h* haploid cells. It is important to liberate the individual ascospores, since mating will otherwise occur between the ascospores within an ascus.

Mixing of the haploid cells allows recombination between more than two lipase genes, enabling "poolwise recombination." Mating brings together new combinations of chimeric genes that can then undergo recombination upon sporulation. The cells are recursively cycled through sporulation, ascospore mixing, and mating until sufficient diversity has been generated by the recursive pairwise recombination of the five lipase genes. The individual chimeric lipase genes either can be screened directly in the haploid yeast cells or transferred to an appropriate expression host.

The process is described above for lipases and yeast; however, any sexual organisms into which genes can be directed can be employed, and any genes, of course, could be substituted for lipases. This process is analogous to the method of stochastic &/or non-stochastic mutagenesis whole genomes by recursive pairwise mating. The diversity, however, in the whole genome case is distributed throughout the host genome rather than localized to specific loci.

#### **8.15.5 USING YACs TO CLONE UNLINKED GENES BUT FUNCTIONALLY IMPORTANT GENES FROM ONE SPECIES INTO ANOTHER**

Stochastic &/or non-stochastic mutagenesis of YACs is particularly amenable to transfer of unlinked but functionally related genes from one species to another, particularly where such genes have not been identified. Such is the case for several commercially important natural products, such as taxol. Transfer of the genes in the metabolic pathway to a different organism is often desirable because organisms naturally producing such compounds are not well suited for mass culturing.

Clusters of such genes can be isolated by cloning a total genomic library of DNA from an organisms producing a useful compound into a YAC library. The YAC library is then transformed into yeast. The yeast is sporulated and mated such that recombination occurs between YACs and/or between YACs and natural yeast chromosomes.

Selection/screening is then performed for expression of the desired collection of genes. If the genes encode a biosynthetic pathway, expression can be detected from the appearance of product of the pathway. Production of individual enzymes in the pathway, or intermediates of the final expression product or capacity of cells to metabolize such intermediates indicates partial acquisition of the synthetic pathway. The original library or a different library can be introduced into cells surviving/selection screening, and further rounds of recombination and selection/screening can be performed until the end product of the desired metabolic pathway is produced.

#### **8.15.6 YAC-YAC STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

If a phenotype of interest can be isolated to a single stretch of genomic DNA less than 2 megabases in length, it can be cloned into a YAC and replicated in *S. cerevisiae*. The cloning of similar stretches of DNA from related hosts into an identical YAC results in a population of yeast cells each harboring a YAC having a homologous insert effecting a desired phenotype. The recursive breeding of these yeast cells allows the homologous regions of these YACs to recombine during meiosis, allowing genes, pathways, and clusters to recombine during each cycle of meiosis. After several cycles of mating and segregation, the YAC inserts are well stochastic &/or non-stochastic mutagenized. The now very diverse yeast library could then be screened for phenotypic improvements resulting from the stochastic &/or non-stochastic mutagenesis of the YAC inserts.

#### **8.15.7 YAC-CHROMOSOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS**

"Mitotic" recombination occurs during cell division and results from the recombination of genes during replication. This type of recombination is not limited to that between sister chromatids and can be enhanced by agents that induce recombination machinery, such as nicking chemicals and ultraviolet irradiation. Since it is often difficult to directly mate across a species barrier, it is possible to induce the recombination of

homologous genes originating from different species by providing the target genes to a desired host organism as a YAC library. The genes harbored in this library are then induced to recombine with homologous genes on the host chromosome by enhanced mitotic recombination. This process is carried out recursively to generate a library of diverse organisms and then screened for those having the desired phenotypic improvements. The improved subpopulation is then mated recursively as above to identify new strains having accumulated multiple useful genetic alterations.

#### **8.15.8 ACCUMULATION OF MULTIPLE YACs HARBORING USEFUL GENES**

The accumulation of multiple unlinked genes that are required for the acquisition or improvement of a given phenotype can be accomplished by the stochastic &/or non-stochastic mutagenesis of YAC libraries. Genomic DNA from organisms having desired phenotypes, such as ethanol tolerance, thermotolerance, and the ability to ferment pentose sugars are pooled, fragmented and cloned into several different YAC vectors, each having a different selective marker (his, ura, ade, etc). *S. cerevisiae* are transformed with these libraries, and selected for their presence (using selective media i.e uracil dropout media for the YAC containing the ura3 selective marker) and then screened for having acquired or improved a desired phenotype.

Surviving cells are pooled, mated recursively, and selected for the accumulation of multiple YACs (by propagation in medium with multiple nutritional dropouts). Cells that acquire multiple YACs harboring useful genomic inserts are identified by further screening. Optimized strains can be used directly, however, due to the burden a YAC may pose to a cell, the relevant YAC inserts can be minimized, subcloned, and recombined into the host chromosome, to generate a more stable production strain.



### 8.15.9 CHOICE OF HOST SSF ORGANISM

One example use for the present invention is to create an improved yeast for the production of ethanol from lignocellulosic biomass. Specifically, a yeast strain with improved ethanol tolerance and thermostability/thermotolerance is desirable. Parent yeast strains known for good behavior in a Simultaneous Saccharification and Fermentation (SSF) process are identified. These strains are combined with others known to possess ethanol," tolerance and/or thermostability.

*S. cerevisiae* is highly amenable to development for optimized SSF processes. It inherently possesses several traits for this use, including the ability to import and ferment a variety of sugars such as sucrose, glucose, galactose, maltose and maltotriose. Also, yeast has the capability to flocculate, enabling recovery of the yeast biomass at the end of a fermentation cycle, and allowing its re-use in subsequent bioprocesses. This is an important property in that it optimizes the use of nutrients in the growth medium. *S. cerevisiae* is also highly amenable to laboratory manipulation, has highly characterized genetics and possesses a sexual reproductive cycle. *S. cerevisiae* may be grown under either aerobic or anaerobic conditions, in contrast to some other potential SSF organisms that are strict anaerobes (e.g. *Clostridium* spp.), making them very difficult to handle in the laboratory. *S. cerevisiae* are also "generally regarded as safe" ("GRAS"), and, due to its widespread use for the production of important comestibles for the general public (e.g. beer, wine, bread, etc), is generally familiar and well known. *S. cerevisiae* is commonly used in fermentative processes, and the familiarity in its handling by fermentation experts eases the introduction of novel improved yeast strains into the industrial setting. *S. cerevisiae* strains that previously have been identified as particularly good SSF organisms, for example, *S. cerevisiae* D<sub>5</sub>A (ATCC200062) (South CR and Lynd LR. (1994) Appl. Biochem. Biotechnol. 45/46: 467-481; Ranatunga TD et al. (1997) Biotechnol. Lett. 19:1125-1127) can be used for starting materials. In addition, other industrially used *S. cerevisiae* strains are optionally used as host strains, particularly those showing desirable fermentative characteristics, such as *S. cerevisiae* Y567 (ATCC24858) (Sitton OC et al. (1979) Process Biochem. 14(9): 7-10; Sitton OC et al. (1981) Adv. Biotechnol. 2: 231-237; McMurrough I et al. (1971) Folia Microbiol. 16: 346-349) and *S. cerevisiae* ACA 174 (ATCC 60868) (Benitez T et al. (1983) Appl. Environ. Microbiol. 45: 1429-1436;

Chem. Eng. J. 50: B I 7-B22, 1992), which have been shown to have desirable traits for large-scale fermentation.

#### 8.15.10 CHOICE OF ETHANOL TOLERANT STRAINS

Many strains of *S. cerevisiae* have been isolated from high-ethanol environments, and have survived in the ethanol-rich environment by adaptive evolution. For example, strains from Sherry wine aging ("Flor" strains) have evolved highly functional mitochondria to enable their survival in a high-ethanol environment. It has been shown that transfer of these wine yeast mitochondria to other strains increases the recipient's resistance to high ethanol concentration, as well as thermotolerance (Jimenez, J. and Benitez, T (1988) Curr. Genet. 13: 461-469). There are several flor strains deposited in the ATCC, for example *S. cerevisiae* MY91 (ATCC 201301), MY138 (ATCC 201302), C5 (ATCC 201298), ET7 (ATCC 201299), LA6 (ATCC 201300), OSB21 (ATCC 201303), F23 (*S. globosus* ATCC 90920). Also, several flor strains of *S. uvarum* and *Torulaspora pretoriensis* have been deposited. Other ethanol-tolerant wine strains include *S. cerevisiae* ACA 174 (ATCC 60868), 15% ethanol, and *S. cerevisiae* A54 (ATCC 90921), isolated from wine containing 18% (v/v) ethanol, and NRCC 202036 (ATCC 46534), also a wine yeast. Other *S. cerevisiae* ethanolologens that additionally exhibit enhanced ethanol tolerance include ATCC 24858, ATCC 24858, G 3706 (ATCC 42594), NRRL Y-265 (ATCC 60593), and ATCC 24845 - ATCC 24860. A strain of *S. pastorianus* (*S. carlsbergensis* ATCC 2345) has high ethanol-tolerance (13% v/v). *S. cerevisiae* Sa28 (ATCC 26603), from Jamaican cane juice sample, produces high levels of alcohol from molasses, is sugar tolerant, and produces ethanol from wood acid hydrolyzate. Several of the listed strains, as well as additional strains can be used as starting materials for breeding ethanol tolerance.

#### 8.15.11 CHOICE OF TEMPERATURE TOLERANT STRAINS

A few temperature tolerant strains have been reported, including the highly flocculent strain *S. pastorianus* SA 23 (*S. carlsbergensis* ATCC 26602), which produces

ethanol at elevated temperatures, and *S. cerevisiae* Kyokai 7 (*S. sake*, ATCC 26422), a sake yeast tolerant to brief heat and oxidative stress. Ballesteros et al ((1991) Appl. Biochem. Biotechnol. 28/29: 307-315) examined 27 strains of yeast for their ability to grow and ferment glucose in the 32-45°C temperature range, including *Saccharomyces*, *Kluyveromyces* and *Candida* spp. Of these, the best thermotolerant clones were *Kluyveromyces marxianus* LG and *Kluyveromyces fragilis* 2671 (Ballesteros et al (1993) Appl. Biochem. Biotechnol. 39/40: 201-211). *S. cerevisiae*-pretoriensis FDHII was somewhat thermotolerant, however was poor in ethanol tolerance. Recursive recombination of this strain with others that display ethanol tolerance can be used to acquire the thermotolerant characteristics of the strain in progeny which also display ethanol tolerance. *Candida acidothermophilum* (Issatchenkia orientalis, ATCC 20381) is a good SSF strain that also exhibits improved performance in ethanol production from lignocellulosic biomass at higher SSF temperatures than *S. cerevisiae* D5A (Kadam, KL, Schmidt, SL (1997) Appl. Microbial. Biotechnol. 48: 709-713). This strain can also be a genetic contributor to an improved SSF strain.

#### 8.15.12 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF STRAINS

In those instances where strains are highly related, a recursive mating strategy may be pursued. For example, a population of haploid *S. cerevisiae* (a and ) are mutagenized and screened for improved EtOH or thermal tolerance. The improved haploid subpopulation are mixed together and mated as a pool and induced to sporulate. The resulting haploid spores are freed by degrading the asci wall and mixed. The freed spores are then induced to mate and sporulate recursively. This process is repeated a sufficient number of times to generate all possible mutant combinations. The whole genome stochastic &/or non-stochastic mutagenized population (haploid) is then screened for further EtOH or thermal tolerance.

When strains are not sufficiently related for recursive mating, formats based on protoplast fusion may be employed. Recursive and poolwise protoplast fusion can be performed to generate chimeric populations of diverse parental strains. The resultant pool

of progeny is selected and screened to identify improved ethanol and thermal tolerant strains.

Alternatively, a YAC-based Whole Genome Stochastic &/or non-stochastic mutagenesis format can be used. In this format, YACs are used to shuttle large chromosomal fragments between strains. As detailed earlier, recombination occurs between YACs or between YACs, and the host chromosomes. Genomic DNA from organisms having desired phenotypes are pooled, fragmented and cloned into several different YAC vectors, each having a different selective marker (his, ura, ade, etc). *S. cerevisiae* are transformed with these libraries, and selected for their presence (using selective media, i.e. uracil dropout media for the YAC containing the Ura 3 selective marker) and then screened for having acquired or improved a desired phenotype. Surviving cells are pooled, mated recursively (as above), and selected for the accumulation of multiple YACs (by propagation in medium with multiple nutritional dropouts). Cells that acquire multiple YACs harboring useful genomic inserts are identified by further screening (see below).

#### **8.15.13 SELECTION FOR IMPROVED STRAINS**

Having produced large libraries of novel strains by mutagenesis and recombination, a first task is to isolate those strains that possess improvements in the desired phenotypes. Identification of the organism libraries is facilitated where the desired key traits are selectable phenotypes. For example, ethanol has different effects on the growth rate of a yeast population, viability, and fermentation rate. Inhibition of cell growth and viability increases with ethanol concentration, but high fermentative capacity is only inhibited at higher ethanol concentrations. Hence, selection of growing cells in ethanol is a viable approach to isolate ethanol-tolerant strains. Subsequently, the selected strains may be analyzed for their fermentative capacity to produce ethanol. Provided that growth and media conditions are the same for all strains (parents and progeny), a hierarchy of ethanol tolerance may be constructed.

Simple selection schemes for identification of thermal tolerant and ethanol tolerant strains are available and, in this case, are based on those previously designed to identify potentially useful SSF strains. Selection of ethanol tolerance is performed by exposing the population to ethanol, then plating the population and looking for growth. Colonies capable of growing after exposure to ethanol can be re-exposed to a higher concentration of ethanol and the cycle repeated until the most tolerant strains are selected. In order to discern strains possessing heritable ethanol tolerance from with temporarily acquired adaptations, these cycles may be punctuated with cycles of growth in the absence of selection (e.g. no ethanol).

Alternatively, the mixed population can be grown directly at increasing concentrations of ethanol, and the most tolerant strains enriched (Aguilera and Benitez, 1986, Arch Microbiol 4:337-44). For example this enrichment could be carried out in a chemostat or turbidostat. Similar selections can be developed for thermal tolerance, in which strains are identified by their ability to grow after a heat treatment, or directly for growth at elevated temperatures (Ballesteros et al., 1991, Applied Biochem and Biotech, 28:307-315). The best strains identified by these selections will be assayed more thoroughly in subsequent screens for ethanol, thermal tolerance or other properties of interest.

In one aspect, organisms having increased ethanol tolerance are selected for. A population of natural *S. cerevisiae* isolates are mutagenized. This population is then grown under fermentor conditions under low initial ethanol concentrations. Once the culture has reached saturation, the culture is diluted into fresh medium having a slightly higher ethanol content. This process of successive dilution into medium of incrementally increasing ethanol concentration is continued until a threshold of ethanol tolerance is reached. The surviving mutant population having the highest ethanol tolerance are then pooled and their genomes recombined by any method noted herein. Enrichment could also be achieved by a continuous culture in a chemostat or turbidostat in which temperature or ethanol concentrations are progressively elevated. The resulting stochastic &/or non-stochastic mutagenized population are then exposed once again to the enrichment strategy

but at a higher starting medium ethanol concentration. This strategy is optionally applied for the enrichment of thermotolerant cells and for the enrichment of cells having combined thermo- and ethanol tolerance.

#### **8.15.14 SCREENING FOR IMPROVED STRAINS**

Strains showing viability in initial selections are assayed more quantitatively for improvements in the desired properties before being restochastic &/or non-stochastic mutagenized with other strains.

Progeny resulting from mutagenesis of a strain, or those pre-selected for their ethanol tolerance and/or thermostability, can be plated on non-selective agar. Colonies can be picked robotically into microtiter dishes and grown. Cultures are replicated to fresh microtiter plates, and the replicates are incubated under the appropriate stress condition(s). The growth or metabolic activity of individual clones may be monitored and ranked. Indicators of viability can range from the size of growing colonies on solid media, density of growing cultures, or color change of a metabolic activity indicator added to liquid media. Strains that show the greatest viability are then mixed and stochastic &/or non-stochastic mutagenized, and the resulting progeny are rescreened under more stringent conditions

#### **8.15.15 DEVELOPMENT OF A YEAST STRAIN CAPABLE OF CONVERTING CELLULOSE TO MONOMERIC SUGARS**

Once a strain of yeast exhibiting thermotolerance and ethanol tolerance is developed, the degradation of cellulose to monomeric sugars is provided by the inclusion to the host strain of an efficient cellulase degradation pathway.

Additional desirable characteristic can be useful to enhance the production of ethanol by the host. For example, inclusion of heterologous enzymes and pathways that broaden the substrate sugar range may be performed. "Tuning" of the strain can be accomplished by the addition of various other traits, or the restoration of certain endogenous traits that are desirable, but lost during the recombination procedures.

### 8.15.16 CONFERRING OF CELLULASE ACTIVITY

A vast number of cellulases and cellulase degradation systems have been characterized from fungi, bacteria and yeast (see reviews by Beguin, P and Aubert, J-P (1994) FEMS Microbial. Rev. 13: 25-58; Ohima, K. et al. (1997) Biotechnol. Genet. Eng. Rev. 14: 365414). An enzymatic pathway required for efficient saccharification of cellulose involves the synergistic action of endoglucanases (endo-1,4- $\beta$ -D-glucanases, EC 3.2.1.4), exocellobiohydrolases (exo-1,4- $\beta$ -D-glucanases, EC 3.2.1.91), and  $\beta$ -glucosidases (cellobiases, 1,4- $\beta$ -D-glucanases EC 3.2.1.21). The heterologous production of cellulase enzymes in the ethanologen would enable the saccharification of cellulose, producing monomeric sugars that may be used by the organism for ethanol production. There are several advantages to the heterologous expression of a functional cellulase pathway in the ethanologen. For example, the SSF process would eliminate the need for a separate bioprocess step for saccharification, and would ameliorate end-product inhibition of cellulase enzymes by accumulated intermediate and product sugars.

Naturally occurring cellulase pathways are inserted into the ethanologen, or one may choose to use custom improved "hybrid" cellulase pathways, employing the coordinate action of cellulases derived from different natural sources, including thermophiles.

Several cellulases from non-Saccharomyces have been produced and secreted from this organism successfully, including bacterial, fungal, and yeast enzymes, for example T. reesei CBH I ((Shoemaker (1994), in "The Cellulase System of Trichoderma reesei: Trichoderma strain improvement and Expression of Trichoderma cellulases in Yeast," Online, Pinner, UK, 593-600). It is possible to employ straightforward metabolic engineering techniques to engender cellulase activity in Saccharomyces. Also, yeast have been forced to acquire elements of cellulose degradation pathways by protoplast fusion (e.g. intergeneric hybrids of Saccharomyces cerevisiae and Zygosaccharomyces fermentati, a cellobiase-producing yeast, have been created (Pina A, et. al. (1986) Appl. Environ. Microbial. 51: 995- 1003). In general, any cellulase component enzyme that

derives from a closely related yeast organism could be transferred by protoplast fusion. Cellobiases produced by a somewhat broader range of yeast may be accessed by whole genome stochastic &/or non-stochastic mutagenesis in one of its many formats (e.g. whole, fragmented, YAC-based).

Optimally, the cellulase enzymes to be used should exhibit good synergy, an appropriate level of expression and secretion from the host, good specific activity (i.e. resistance to host degradation factors and enzyme modification) and stability in the desired SSF environment. An example of a hybrid cellulose degradation pathway having excellent synergy includes the following enzymes: CBH I exocellobiohydrolase of *Trichoderma reesei*, the *Acidothermus cellulolyticus* E1 endoglucanase, and the *Thermomonospora fusca* E3 exocellulase (Baker, et. al. (1998) Appl. Biochem. Biotechnol. 70-72: 395-403).

It is suggested here that these enzymes (or improved mutants thereof) be considered for use in the SSF organism, along with a cellobiase ( - g l ucosidase), such as that from *Candida peltata*. Other possible cellulase systems to be considered should possess particularly good activity against crystalline cellulose, such as the *T. reesei* cellulase system (Teeri, TT, et. al. (1998) Biochem. Soc. Trans. 26: 173-178), or possess particularly good thermostability characteristics (e.g. cellulase systems from thermophilic organisms, such as *Thermomonospora fusca* (Zhang, S., et. al. (1995) Biochem.. 34: 3 3 86-3 3 5).

A rational approach to the cloning of cellulases in the ethanologenic yeast host could be used. For example, known cellulase genes are cloned into expression cassettes utilizing *S. cerevisiae* promoter sequences, and the resultant linear fragments of DNA may be transformed into the recipient host by placing short yeast sequences at the termini to encourage site-specific integration into the genome. This is preferred to plasmidic transformation for reasons of genetic stability and maintenance of the transforming DNA.

If an entire cellulose degradative pathway were introduced, a selection could be implemented in an agar-plate-based format, and a large number of clones could be assayed



for cellulase activity in a short period of time. For example, selection for an exocellulase may be accessible by providing a soluble oligocellulose substrate or carboxymethylcellulose (CMC) as a sole carbon source to the host, otherwise unable to grow on agar containing this sole carbon source. Clones producing active cellulase pathways would grow by virtue of their ability to produce glucose.

Alternatively, if the different cellulases were to be introduced sequentially, it would be useful to first introduce a cellobiase, enabling a selection using commercially available cellobiose as a sole carbon source. Several strains of *S. cerevisiae* that are able to grow on cellobiose have been created by introduction of a cellobiase gene (e.g. Rajoka MI, et. al. (1998) *Folia Microbiol. (Praha)* 43, 129-135; Skory, CD, et. al. (1996) *Curr. Genet.* 30, 417-422; D'Auria, S, et. al. (1996) *Appl. Biochem. Biotechnol.* 61, 157-166; Adam, AC, et. al. (1995) *Yeast* 11, 395-406; Adam, AC (1991) *Curr. Genet.* 20, 5-8).

Subsequent transformation of this organism with CBHI exocellulase can be selected for by growth on a cellulose substrate such as carboxymethylcellulose (CMC). Finally, addition of an endoglucanase creates a yeast strain with improved crystalline degradation capacity.

#### 8.15.17 CONFERRING OF PENTOSE SUGAR UTILIZATION

Inclusion of pentose sugar utilization pathways is an important facet to a potentially useful SSF organism. The successful expression of xylose sugar utilization pathways for ethanol production has been reported in *Saccharomyces* (e.g. Chen, ZD and Ho, NWY (1993) *Appl. Biochem. Biotechnol.* 39/40 135-147).

It would also be useful to accomplish L-arabinose substrate utilization for ethanol production in the *Saccharomyces* host. Yeast strains that utilize L-arabinose include some *Candida* and *Pichia* spp. (McMillan JD and Boynton BL (1994) *Appl. Biochem. Biotechnol.* 45-46: 569-584; Dien BS, et al. (1996) *Appl. Biochem. Biotechnol.* 57-58: 233-242). Genes necessary for arabinose fermentation in *E. coli* could also be introduced

by rational means (e.g. as has been performed previously in *Z. mobilis* (Deanda K, et. al. (1996) Appl. Environ. Microbial. 62: 4465-4470))

#### 8.15.18 CONFERRING OF OTHER USEFUL ACTIVITIES

Several other traits that are important for optimization of an SSF strain have been shown to be transferable to *S. cerevisiae*. Like thermal tolerance, cellulase activity and pentose sugar utilization, these traits may not normally be exhibited by *Saccharomyces* (or the particular strain of *Saccharomyces* being used as a host), and may be added by genetic means.

For example, expression of human muscle acylphosphatase in *S. cerevisiae* has been suggested to increase ethanol production (Rougei, G., et. al. (1996) Biotechnol. App. Biochem. 23: 273- 278).

It can occur that evolved stress-tolerant SSF strain acquire some undesirable mutations in the course of the evolution strategy. Indeed, this is a pervasive problem in strain improvement strategies that rely on mutagenesis techniques, and can result in highly unstable or fragile production strains. It is possible to restore some of these desirable traits by rational methods such as cloning of specific genes that have been knocked out or negatively influenced in the previous rounds of strain improvement. The advantage to this approach is specificity- the offending gene may be targeted directly. The disadvantage is that it may be time-consuming and repetitious if several genes have been compromised, and it only addresses problems that have been characterized. A preferred (and more traditional) approach to the removal of undesirable/deleterious mutations is to back-cross the evolved strain to a desirable parent strain (e.g. the original "host" SSF strain). This strategy has been employed successfully throughout strain improvement where accessible (i.e. for organisms that have sexual cycles of reproduction). When lacking the advantage of a sexual process, it has been accomplished by using other methods, such as parasexual recombination or protoplast fusion.

For example, the ability to flocculate was conferred on a non- flocculating strain of *S. cerevisiae* by protoplast fusion with a flocculation competent *S. cerevisiae* (Watari, J., et. al (1990) Agric. Biol. Chem. 54: 1677-1681).

## 8.16 METHOD OF IN VIVO AND IN VITRO DNA SHUFFLING

### 8.16.1 Applications

Disclosed is a method of producing random polynucleotides by introducing two or more related polynucleotides into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment. Also provided are vector and expression vehicles including such polynucleotides, polypeptides expressed by the hybrid polynucleotides and a method for screening for hybrid polypeptides

### 8.16.2 Experimental Applications

This invention relates generally to recombination and more specifically to a method for preparing polynucleotides encoding a polypeptide by a method of *in vivo* re-assortment of polynucleotide sequences containing regions of partial homology, assembling the polynucleotides to form at least one polynucleotide and screening the polynucleotides for the production of polypeptide(s) having a useful property.

### 8.16.3 History

An exceedingly large number of possibilities exist for purposeful and random combinations of amino acids within a protein to produce useful hybrid proteins and their corresponding biological molecules encoding for these hybrid proteins, *i.e.*, DNA, RNA. Accordingly, there is a need to produce and screen a wide variety of such hybrid proteins for a useful utility, particularly widely varying random proteins.

The complexity of an active sequence of a biological macromolecule (*e.g.*, proteins, DNA) has been called its information content ("IC"), which has been defined as the resistance of the active protein to amino acid sequence variation (calculated from the minimum number of

invariable amino acids (bits) required to describe a family of related sequences with the same function. Proteins that are more sensitive to random mutagenesis have a high information content.

Molecular biology developments, such as molecular libraries, have allowed the identification of quite a large number of variable bases, and even provide ways to select functional sequences from random libraries. In such libraries, most residues can be varied (although typically not all at the same time) depending on compensating changes in the context. Thus, while a 100 amino acid protein can contain only 2,000 different mutations,  $20^{100}$  sequence combinations are possible.

Information density is the IC per unit length of a sequence. Active sites of enzymes tend to have a high information density. By contrast, flexible linkers of information in enzymes have a low information density.

Current methods in widespread use for creating alternative proteins in a library format are error-prone polymerase chain reactions and cassette mutagenesis, in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide. In both cases, a substantial number of mutant sites are generated around certain sites in the original sequence.

#### **8.16.3.1 Error-prone PCR**

Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. In a mixture of fragments of unknown sequence, error-prone PCR can be used to mutagenize the mixture. The published error-prone PCR protocols suffer from a low processivity of the polymerase. Therefore, the protocol is unable to result in the random mutagenesis of an average-sized gene. This inability limits the practical application of error-prone PCR. Some computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the large-scale block changes that are required for continued and dramatic sequence evolution. Further, the published error-prone PCR protocols do not allow for amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical

application. In addition, repeated cycles of error-prone PCR can lead to an accumulation of neutral mutations with undesired results, such as affecting a protein's immunogenicity but not its binding affinity.

#### **8.16.3.2 Oligonucleotide-Directed Mutagenesis**

In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping them into families, arbitrarily choosing a single family, and reducing it to a consensus motif. Such motif is resynthesized and reinserted into a single gene followed by additional selection. This step process constitutes a statistical bottleneck, is labor intensive, and is not practical for many rounds of mutagenesis.

Error-prone PCR and oligonucleotide-directed mutagenesis are thus useful for single cycles of sequence fine tuning, but rapidly become too limiting when they are applied for multiple cycles.

Another limitation of error-prone PCR is that the rate of down-mutations grows with the information content of the sequence. As the information content, library size, and mutagenesis rate increase, the balance of down-mutations to up-mutations will statistically prevent the selection of further improvements (statistical ceiling).

#### **8.16.3.3 Cassette Mutagenesis**

In cassette mutagenesis, a sequence block of a single template is typically replaced by a (partially) randomized sequence. Therefore, the maximum information content that can be obtained is statistically limited by the number of random sequences (*i.e.*, library

size). This eliminates other sequence families which are not currently best, but which may have greater long term potential.

Also, mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round. Thus, such an approach is tedious and impractical for many rounds of mutagenesis.

Thus, error-prone PCR and cassette mutagenesis are best suited, and have been widely used, for fine-tuning areas of comparatively low information content. One apparent exception is the selection of an RNA ligase ribozyme from a random library using many rounds of amplification by error-prone PCR and selection.

In nature, the evolution of most organisms occurs by natural selection and sexual reproduction. Sexual reproduction ensures mixing and combining of the genes in the offspring of the selected individuals. During meiosis, homologous chromosomes from the parents line up with one another and cross-over part way along their length, thus randomly swapping genetic material. Such swapping or shuffling of the DNA allows organisms to evolve more rapidly.

In recombination, because the inserted sequences were of proven utility in a homologous environment, the inserted sequences are likely to still have substantial information content once they are inserted into the new sequence.

#### **8.16.3.4 Applied Molecular Evolution**

The term Applied Molecular Evolution ("AME") means the application of an evolutionary design algorithm to a specific, useful goal. While many different library formats for AME have been reported for polynucleotides, peptides and proteins (phage, lacI and polysomes), none of these formats have provided for recombination by random cross-overs to deliberately create a combinatorial library.

Theoretically there are 2,000 different single mutants of a 100 amino acid protein. However, a protein of 100 amino acids has  $20^{100}$  possible sequence combinations, a number which is too large to exhaustively explore by conventional methods. It would be

advantageous to develop a system which would allow generation and screening of all of these possible combination mutations.

#### 8.16.3.5 Reported *in vivo* Recombination Systems

Some workers in the art have utilized an *in vivo* site specific recombination system to generate hybrids of combine light chain antibody genes with heavy chain antibody genes for expression in a phage system. However, their system relies on specific sites of recombination and is limited accordingly. Simultaneous mutagenesis of antibody CDR regions in single chain antibodies (scFv) by overlapping extension and PCR have been reported.

Others have described a method for generating a large population of multiple hybrids using random *in vivo* recombination. This method requires the recombination of two different libraries of plasmids, each library having a different selectable marker. The method is limited to a finite number of recombinations equal to the number of selectable markers existing, and produces a concomitant linear increase in the number of marker genes linked to the selected sequence(s).

*In vivo* recombination between two homologous, but truncated, insect-toxin genes on a plasmid has been reported as a method of producing a hybrid gene. The *in vivo* recombination of substantially mismatched DNA sequences in a host cell having defective mismatch repair enzymes, resulting in hybrid molecule formation has been reported.

#### 8.16.4 Strategies

In one aspect this invention provides a method that utilizes the natural property of cells to recombine molecules and/or to mediate reductive processes that reduce the complexity of sequences and extent of repeated or consecutive sequences possessing regions of homology.

It is an object of the present invention to provide a method for generating hybrid polynucleotides encoding biologically active hybrid polypeptides with enhanced activities. In accomplishing these and other objects, there has been provided, in accordance with one aspect of the invention, a method for introducing polynucleotides into a suitable host cell and growing the host cell under conditions which produce a hybrid polynucleotide.

In another aspect of the invention, the invention provides a method for screening for biologically active hybrid polypeptides encoded by hybrid polynucleotides. The present method allows for the identification of biologically active hybrid polypeptides with enhanced biological activities.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

#### 8.16.5 Possible Uses

The invention described herein is directed to the use of repeated cycles of reductive reassortment, recombination and selection which allow for the directed molecular evolution of highly complex linear sequences, such as DNA, RNA or proteins thorough recombination.



*In vivo* shuffling of molecules can be performed utilizing the natural property of cells to recombine multimers. While recombination *in vivo* has provided the major natural route to molecular diversity, genetic recombination remains a relatively complex process that involves 1) the recognition of homologies; 2) strand cleavage, strand invasion, and metabolic steps leading to the production of recombinant chiasma; and finally 3) the resolution of chiasma into discrete recombined molecules. The formation of the chiasma requires the recognition of homologous sequences.

#### **8.16.5.1 Production of a Hybrid Polynucleotide**

In a preferred embodiment, the invention relates to a method for producing a hybrid polynucleotide from at least a first polynucleotide and a second polynucleotide. The present invention can be used to produce a hybrid polynucleotide by introducing at least a first polynucleotide and a second polynucleotide which share at least one region of partial sequence homology into a suitable host cell. The regions of partial sequence homology promote processes which result in sequence reorganization producing a hybrid polynucleotide. The term "hybrid polynucleotide", as used herein, is any nucleotide sequence which results from the method of the present invention and contains sequence from at least two original polynucleotide sequences. Such hybrid polynucleotides can result from intermolecular recombination events which promote sequence integration between DNA molecules. In addition, such hybrid polynucleotides can result from intramolecular reductive reassortment processes which utilize repeated sequences to alter a nucleotide sequence within a DNA molecule.

The invention provides a means for generating hybrid polynucleotides which may encode biologically active hybrid polypeptides. In one aspect, the original polynucleotides encode biologically active polypeptides. The method of the invention produces new hybrid polypeptides by utilizing cellular processes which integrate the sequence of the original polynucleotides such that the resulting hybrid polynucleotide encodes a polypeptide demonstrating activities derived from the original biologically active polypeptides. For example, the original polynucleotides may encode a particular enzyme from different microorganisms. An enzyme encoded by a first polynucleotide from one

organism may, for example, function effectively under a particular environmental condition, *e.g.* high salinity. An enzyme encoded by a second polynucleotide from a different organism may function effectively under a different environmental condition, such as extremely high temperatures. A hybrid polynucleotide containing sequences from the first and second original polynucleotides may encode an enzyme which exhibits characteristics of both enzymes encoded by the original polynucleotides. Thus, the enzyme encoded by the hybrid polynucleotide may function effectively under environmental conditions shared by each of the enzymes encoded by the first and second polynucleotides, *e.g.*, high salinity and extreme temperatures.

#### 8.16.5.1.1 Encoded Enzymes

Enzymes encoded by the original polynucleotides of the invention include, but are not limited to; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. A hybrid polypeptide resulting from the method of the invention may exhibit specialized enzyme activity not displayed in the original enzymes. For example, following recombination and/or reductive reassortment of polynucleotides encoding hydrolase activities, the resulting hybrid polypeptide encoded by a hybrid polynucleotide can be screened for specialized hydrolase activities obtained from each of the original enzymes, *i.e.* the type of bond on which the hydrolase acts and the temperature at which the hydrolase functions. Thus, for example, the hydrolase may be screened to ascertain those chemical functionalities which distinguish the hybrid hydrolase from the original hydrolyases, such as: (a) amide (peptide bonds), *i.e.* proteases; (b) ester bonds, *i.e.* esterases and lipases; (c) acetals, *i.e.*, glycosidases and, for example, the temperature, pH or salt concentration at which the hybrid polypeptide functions.

#### 8.16.5.1.2 Sources Of The Original Polynucleotides

Sources of the original polynucleotides may be isolated from individual organisms ("isolates"), collections of organisms that have been grown in defined media ("enrichment cultures"), or, most preferably, uncultivated organisms ("environmental samples"). The

use of a culture-independent approach to derive polynucleotides encoding novel bioactivities from environmental samples is most preferable since it allows one to access untapped resources of biodiversity.

"Environmental libraries" are generated from environmental samples and represent the collective genomes of naturally occurring organisms archived in cloning vectors that can be propagated in suitable prokaryotic hosts. Because the cloned DNA is initially extracted directly from environmental samples, the libraries are not limited to the small fraction of prokaryotes that can be grown in pure culture. Additionally, a normalization of the environmental DNA present in these samples could allow more equal representation of the DNA from all of the species present in the original sample. This can dramatically increase the efficiency of finding interesting genes from minor constituents of the sample which may be under-represented by several orders of magnitude compared to the dominant species.

For example, gene libraries generated from one or more uncultivated microorganisms are screened for an activity of interest. Potential pathways encoding bioactive molecules of interest are first captured in prokaryotic cells in the form of gene expression libraries. Polynucleotides encoding activities of interest are isolated from such libraries and introduced into a host cell. The host cell is grown under conditions which promote recombination and/or reductive reassortment creating potentially active biomolecules with novel or enhanced activities.

The microorganisms from which the polynucleotide may be prepared include prokaryotic microorganisms, such as Eubacteria and Archaeobacteria, and lower eukaryotic microorganisms such as fungi, some algae and protozoa. Polynucleotides may be isolated from environmental samples in which case the nucleic acid may be recovered without culturing of an organism or recovered from one or more cultured organisms. In one aspect, such microorganisms may be extremophiles, such as hyperthermophiles, psychrophiles, psychrotrophs, halophiles, barophiles and acidophiles. Polynucleotides encoding enzymes isolated from extremophilic microorganisms are particularly preferred. Such enzymes may function at temperatures above 100°C in terrestrial hot springs and deep sea thermal vents, at temperatures below 0°C in arctic waters, in the saturated salt environment of the Dead

Sea, at pH values around 0 in coal deposits and geothermal sulfur-rich springs, or at pH values greater than 11 in sewage sludge. For example, several esterases and lipases cloned and expressed from extremophilic organisms show high activity throughout a wide range of temperatures and pHs.

#### 8.16.5.1.3 Suitable Host Cells

Polynucleotides selected and isolated as hereinabove described are introduced into a suitable host cell. A suitable host cell is any cell which is capable of promoting recombination and/or reductive reassortment. The selected polynucleotides are preferably already in a vector which includes appropriate control sequences. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or preferably, the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as *E. coli*, *Streptomyces*, *Salmonella typhimurium*; fungal cells, such as yeast; insect cells such as *Drosophila S2* and *Spodoptera Sf9*; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; and plant cells. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

##### 8.16.5.1.3.1 Mammalian Cell Culture Systems

With particular references to various mammalian cell culture systems that can be employed to express recombinant protein, examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by Gluzman, Cell, 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA

sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Host cells containing the polynucleotides of interest can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan. The clones which are identified as having the specified enzyme activity may then be sequenced to identify the polynucleotide sequence encoding an enzyme having the enhanced activity.

#### 8.16.5.1.4 Generation Of Polynucleotides Encoding Biochemical Pathways

In another aspect, it is envisioned the method of the present invention can be used to generate novel polynucleotides encoding biochemical pathways from one or more operons or gene clusters or portions thereof. For example, bacteria and many eukaryotes have a coordinated mechanism for regulating genes whose products are involved in related processes. The genes are clustered, in structures referred to as "gene clusters," on a single chromosome and are transcribed together under the control of a single regulatory sequence, including a single promoter which initiates transcription of the entire cluster. Thus, a gene cluster is a group of adjacent genes that are either identical or related, usually as to their function. An example of a biochemical pathway encoded by gene clusters are polyketides. Polyketides are molecules which are an extremely rich source of bioactivities, including antibiotics (such as tetracyclines and erythromycin), anti-cancer agents (daunomycin), immunosuppressants (FK506 and rapamycin), and veterinary products (monensin). Many polyketides (produced by polyketide synthases) are valuable as therapeutic agents. Polyketide synthases are multifunctional enzymes that catalyze the biosynthesis of an enormous variety of carbon chains differing in length and patterns of functionality and cyclization. Polyketide synthase genes fall into gene clusters and at least one type (designated type I) of polyketide synthases have large size genes and enzymes, complicating genetic manipulation and *in vitro* studies of these genes/proteins.

The ability to select and combine desired components from a library of polyketides, or fragments thereof, and postpolyketide biosynthesis genes for generation of

novel polyketides for study is appealing. The method of the present invention makes it possible to facilitate the production of novel polyketide synthases through intermolecular recombination.

#### 8.16.5.1.5 Gene Cluster DNA

Preferably, gene cluster DNA can be isolated from different organisms and ligated into vectors, particularly vectors containing expression regulatory sequences which can control and regulate the production of a detectable protein or protein-related array activity from the ligated gene clusters. Use of vectors which have an exceptionally large capacity for exogenous DNA introduction are particularly appropriate for use with such gene clusters and are described by way of example herein to include the f-factor (or fertility factor) of *E. coli*. This f-factor of *E. coli* is a plasmid which affect high-frequency transfer of itself during conjugation and is ideal to achieve and stably propagate large DNA fragments, such as gene clusters from mixed microbial samples. Once ligated into an appropriate vector, two or more vectors containing different polyketide synthase gene clusters can be introduced into a suitable host cell. Regions of partial sequence homology shared by the gene clusters will promote processes which result in sequence reorganization resulting in a hybrid gene cluster. The novel hybrid gene cluster can then be screened for enhanced activities not found in the original gene clusters.

Therefore, in a preferred embodiment, the present invention relates to a method for producing a biologically active hybrid polypeptide and screening such a polypeptide for enhanced activity by:

introducing at least a first polynucleotide in operable linkage and a second polynucleotide in operable linkage, said at least first polynucleotide and second polynucleotide sharing at least one region of partial sequence homology, into a suitable host cell;

growing the host cell under conditions which promote sequence reorganization resulting in a hybrid polynucleotide in operable linkage;

expressing a hybrid polypeptide encoded by the hybrid polynucleotide;

screening the hybrid polypeptide under conditions which promote identification of enhanced biological activity; and

isolating the a polynucleotide encoding the hybrid polypeptide.

Methods for screening for various enzyme activities are known to those of skill in the art and discussed throughout the present specification. Such methods may be employed when isolating the polypeptides and polynucleotides of the present invention.

The term "isolated" means that material is removed from its original environment (*e.g.*, the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide separated from some or all of the coexisting materials in the natural system, is isolated.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

#### **8.16.5.1.6 Expression Vectors**

As representative examples of expression vectors which may be used there may be mentioned viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral DNA (*e.g.* vaccinia, adenovirus, fowl pox virus, pseudorabies and derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, aspergillus and yeast) Thus, for example, the DNA may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors

include chromosomal, nonchromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used as long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed with the present invention.

A preferred type of vector for use in the present invention contains an f-factor origin replication. The f-factor (or fertility factor) in *E. coli* is a plasmid which effects high frequency transfer of itself during conjugation and less frequent transfer of the bacterial chromosome itself. A particularly preferred embodiment is to use cloning vectors, referred to as "fosmids" or bacterial artificial chromosome (BAC) vectors. These are derived from *E. coli* f-factor which is able to stably integrate large segments of genomic DNA. When integrated with DNA from a mixed uncultured environmental sample, this makes it possible to achieve large genomic fragments in the form of a stable "environmental DNA library."

Another preferred type of vector for use in the present invention is a cosmid vector. Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. Cloning into cosmid vectors is described in detail in Sambrook, *et al.*, Molecular Cloning A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, 1989.

#### 8.16.5.1.6.1 Expression Control Sequence

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct RNA synthesis. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P<sub>R</sub>, P<sub>L</sub> and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector



and promoter is well within the level of ordinary skill in the art. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers.

#### 8.16.5.1.6.2 Selectable Marker Genes

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in *E. coli*.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, *e.g.*, the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), -factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The cloning strategy permits expression via both vector driven and endogenous promoters; vector promotion may be important with expression of genes whose endogenous promoter will not function in *E. coli*.

#### 8.16.5.1.7 Insertion Into a Vector or Plasmid

The DNA isolated or derived from microorganisms can preferably be inserted into a vector or a plasmid prior to probing for selected DNA. Such vectors or plasmids are

preferably those containing expression regulatory sequences, including promoters, enhancers and the like. Such polynucleotides can be part of a vector and/or a composition and still be isolated, in that such vector or composition is not part of its natural environment. Particularly preferred phage or plasmid and methods for introduction and packaging into them are described in detail in the protocol set forth herein.

The selection of the cloning vector depends upon the approach taken, for example, the vector can be any cloning vector with an adequate capacity for multiply repeated copies of a sequence, or multiple sequences that can be successfully transformed and selected in a host cell. One example of such a vector is described in "Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts", Alting-Mecs MA, Short JM, Gene, 1993 Dec 27, 137:1, 93-100. Propagation/maintenance can be by an antibiotic resistance carried by the cloning vector. After a period of growth, the naturally abbreviated molecules are recovered and identified by size fractionation on a gel or column, or amplified directly. The cloning vector utilized may contain a selectable gene that is disrupted by the insertion of the lengthy construct. As reductive reassortment progresses, the number of repeated units is reduced and the interrupted gene is again expressed and hence selection for the processed construct can be applied. The vector may be an expression/selection vector which will allow for the selection of an expressed product possessing desirable biological properties. The insert may be positioned downstream of a functional promoter and the desirable property screened by appropriate means.

#### **8.16.5.1.8 Reductive Reassortment**

*In vivo* reassortment is focused on "inter-molecular" processes collectively referred to as "recombination" which in bacteria, is generally viewed as a "RecA-dependent" phenomenon. The present invention can rely on recombination processes of a host cell to recombine and re-assort sequences, or the cells ability to mediate reductive processes to decrease the complexity of quasi-repeated sequences in the cell by deletion. This process of "reductive reassortment" occurs by an "intra-molecular", RecA-independent process.

Therefore, in another aspect of the present invention, novel polynucleotides can be generated by the process of reductive reassortment. The method involves the generation of constructs containing consecutive sequences (original encoding sequences), their insertion into an appropriate vector, and their subsequent introduction into an appropriate host cell. The reassortment of the individual molecular identities occurs by combinatorial processes between the consecutive sequences in the construct possessing regions of homology, or between quasi-repeated units. The reassortment process recombines and/or reduces the complexity and extent of the repeated sequences, and results in the production of novel molecular species. Various treatments may be applied to enhance the rate of reassortment. These could include treatment with ultra-violet light, or DNA damaging chemicals, and/or the use of host cell lines displaying enhanced levels of "genetic instability". Thus the reassortment process may involve homologous recombination or the natural property of quasi-repeated sequences to direct their own evolution.

#### 8.16.5.1.9 Repeated or "Quasi-Repeated" Sequences

Repeated or "quasi-repeated" sequences play a role in genetic instability. In the present invention, "quasi-repeats" are repeats that are not restricted to their original unit structure. Quasi-repeated units can be presented as an array of sequences in a construct; consecutive units of similar sequences. Once ligated, the junctions between the consecutive sequences become essentially invisible and the quasi-repetitive nature of the resulting construct is now continuous at the molecular level. The deletion process the cell performs to reduce the complexity of the resulting construct operates between the quasi-repeated sequences. The quasi-repeated units provide a practically limitless repertoire of templates upon which slippage events can occur. The constructs containing the quasi-repeats thus effectively provide sufficient molecular elasticity that deletion (and potentially insertion) events can occur virtually anywhere within the quasi-repetitive units.

When the quasi-repeated sequences are all ligated in the same orientation, for instance head to tail or vice versa, the cell cannot distinguish individual units. Consequently, the reductive process can occur throughout the sequences. In contrast, when for example, the units are presented head to head, rather than head to tail, the

inversion delineates the endpoints of the adjacent unit so that deletion formation will favor the loss of discrete units. Thus, it is preferable with the present method that the sequences are in the same orientation. Random orientation of quasi-repeated sequences will result in the loss of reassortment efficiency, while consistent orientation of the sequences will offer the highest efficiency. However, while having fewer of the contiguous sequences in the same orientation decreases the efficiency, it may still provide sufficient elasticity for the effective recovery of novel molecules. Constructs can be made with the quasi-repeated sequences in the same orientation to allow higher efficiency.

#### **8.16.5.1.10 Assembly of Sequences in a Head to Tail Orientation**

Sequences can be assembled in a head to tail orientation using any of a variety of methods, including the following:

- a) Primers that include a poly-A head and poly-T tail which when made single-stranded would provide orientation can be utilized. This is accomplished by having the first few bases of the primers made from RNA and hence easily removed RNaseH.
- b) Primers that include unique restriction cleavage sites can be utilized. Multiple sites, a battery of unique sequences, and repeated synthesis and ligation steps would be required.
- c) The inner few bases of the primer could be thiolated and an exonuclease used to produce properly tailed molecules.

#### **8.16.5.1.11 The Recovery of the Re-assorted Sequences**

The recovery of the re-assorted sequences relies on the identification of cloning vectors with a reduced RI. The re-assorted encoding sequences can then be recovered by amplification. The products are re-cloned and expressed. The recovery of cloning vectors with reduced RI can be effected by:

- 1) The use of vectors only stably maintained when the construct is reduced in complexity.
- 2) The physical recovery of shortened vectors by physical procedures. In this case, the cloning vector would be recovered using standard plasmid isolation procedures and size fractionated on either an agarose gel, or column with a low molecular weight cut off utilizing standard procedures.
- 3) The recovery of vectors containing interrupted genes which can be selected when insert size decreases.
- 4) The use of direct selection techniques with an expression vector and the appropriate selection.

Encoding sequences (for example, genes) from related organisms may demonstrate a high degree of homology and encode quite diverse protein products. These types of sequences are particularly useful in the present invention as quasi-repeats. However, while the examples illustrated below demonstrate the reassortment of nearly identical original encoding sequences (quasi-repeats), this process is not limited to such nearly identical repeats.

The following example demonstrates the method of the invention. Encoding nucleic acid sequences (quasi-repeats) derived from three (3) unique species are depicted. Each sequence encodes a protein with a distinct set of properties. Each of the sequences differs by a single or a few base pairs at a unique position in the sequence which are designated "A", "B" and "C". The quasi-repeated sequences are separately or collectively amplified and ligated into random assemblies such that all possible permutations and combinations are available in the population of ligated molecules. The number of quasi-repeat units can be controlled by the assembly conditions. The average number of quasi-repeated units in a construct is defined as the repetitive index (RI).

Once formed, the constructs may, or may not be size fractionated on an agarose gel according to published protocols, inserted into a cloning vector, and transfected into an

appropriate host cell. The cells are then propagated and "reductive reassortment" is effected. The rate of the reductive reassortment process may be stimulated by the introduction of DNA damage if desired. Whether the reduction in RI is mediated by deletion formation between repeated sequences by an "intra-molecular" mechanism, or mediated by recombination-like events through "inter-molecular" mechanisms is immaterial. The end result is a reassortment of the molecules into all possible combinations.

Optionally, the method comprises the additional step of screening the library members of the shuffled pool to identify individual shuffled library members having the ability to bind or otherwise interact (*e.g.*, such as catalytic antibodies) with a predetermined macromolecule, such as for example a proteinaceous receptor, peptide oligosaccharide, viron, or other predetermined compound or structure.

The displayed polypeptides, antibodies, peptidomimetic antibodies, and variable region sequences that are identified from such libraries can be used for therapeutic, diagnostic, research and related purposes (*e.g.*, catalysts, solutes for increasing osmolarity of an aqueous solution, and the like), and/or can be subjected to one or more additional cycles of shuffling and/or affinity selection. The method can be modified such that the step of selecting for a phenotypic characteristic can be other than of binding affinity for a predetermined molecule (*e.g.*, for catalytic activity, stability oxidation resistance, drug resistance, or detectable phenotype conferred upon a host cell).

#### **8.16.5.1.12 Providing Antibodies Suitable for Affinity Interactions Screening**

The present invention provides a method for generating libraries of displayed antibodies suitable for affinity interactions screening. The method comprises (1) obtaining first a plurality of selected library members comprising a displayed antibody and an associated polynucleotide encoding said displayed antibody, and obtaining said associated polynucleotide encoding for said displayed antibody and obtaining said associated polynucleotides or copies thereof, wherein said associated polynucleotides comprise a region of substantially identical variable region framework sequence, and (2) introducing

said polynucleotides into a suitable host cell and growing the cells under conditions which promote recombination and reductive reassortment resulting in shuffled polynucleotides. CDR combinations comprised by the shuffled pool are not present in the first plurality of selected library members, said shuffled pool composing a library of displayed antibodies comprising CDR permutations and suitable for affinity interaction screening. Optionally, the shuffled pool is subjected to affinity screening to select shuffled library members which bind to a predetermined epitope (antigen) and thereby selecting a plurality of selected shuffled library members. Further, the plurality of selectively shuffled library members can be shuffled and screened iteratively, from 1 to about 1000 cycles or as desired until library members having a desired binding affinity are obtained.

#### **8.16.5.1.13 Introduction of Mutations Into the Original Polynucleotides**

In another aspect of the invention, it is envisioned that prior to or during recombination or reassortment, polynucleotides generated by the method of the present invention can be subjected to agents or processes which promote the introduction of mutations into the original polynucleotides. The introduction of such mutations would increase the diversity of resulting hybrid polynucleotides and polypeptides encoded therefrom. The agents or processes which promote mutagenesis can include, but are not limited to: (+)-CC-1065, or a synthetic analog such as (+)-CC-1065-(N3-Adenine), (see, *Biochem. 31*, 2822-2829 (1992)); a N-acetylated or deacetylated 4'-fluoro-4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see, for example, *Carcinogenesis* vol. 13, No. 5, 751-758 (1992); or a N-acetylated or deacetylated 4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see also, *Id.* 751-758); trivalent chromium, a trivalent chromium salt, a polycyclic aromatic hydrocarbon ("PAH") DNA adduct capable of inhibiting DNA replication, such as 7-bromomethyl-benz[*a*]anthracene ("BMA"), tris(2,3-dibromopropyl)phosphate ("Tris-BP"), 1,2-dibromo-3-chloropropane ("DBCP"), 2-bromoacrolein (2BA), benzo[*a*]pyrene-7,8-dihydrodiol-9-10-epoxide ("BPDE"), a platinum(II) halogen salt, N-hydroxy-2-amino-3-methylimidazo[4,5-*f*]-quinoline ("N-hydroxy-IQ"), and N-hydroxy-2-amino-1-methyl-6-phenylimidazo[4,5-*f*]-pyridine ("N-hydroxy-PhIP"). Especially preferred "means for slowing or halting PCR amplification

consist of UV light (+)-CC-1065 and (+)-CC-1065-(N3-Adenine). Particularly encompassed means are DNA adducts or polynucleotides comprising the DNA adducts from the polynucleotides or polynucleotides pool, which can be released or removed by a process including heating the solution comprising the polynucleotides prior to further processing.

#### **8.16.5.1.14 Production Of Hybrid Or Re-Assorted Polynucleotides**

In another aspect the present invention is directed to a method of producing recombinant proteins having biological activity by treating a sample comprising double-stranded template polynucleotides encoding a wild-type protein under conditions according to the present invention which provide for the production of hybrid or re-assorted polynucleotides.

#### **8.16.5.1.15 Shuffling a Population of Viral Genes of Viral Genomes**

The invention also provides the use of polynucleotide shuffling to shuffle a population of viral genes (*e.g.*, capsid proteins, spike glycoproteins, polymerases, and proteases) or viral genomes (*e.g.*, paramyxoviridae, orthomyxoviridae, herpesviruses, retroviruses, reoviruses and rhinoviruses). In an embodiment, the invention provides a method for shuffling sequences encoding all or portions of immunogenic viral proteins to generate novel combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes which are likely to arise in the natural environment as a consequence of viral evolution; (*e.g.*, such as recombination of influenza virus strains).



#### **8.16.5.1.16 Generation of Gene Therapy Vectors and Replication-Defective Gene Therapy Constructs**

The invention also provides a method suitable for shuffling polynucleotide sequences for generating gene therapy vectors and replication-defective gene therapy constructs, such as may be used for human gene therapy, including but not limited to vaccination vectors for DNA-based vaccination, as well as anti-neoplastic gene therapy and other general therapy formats.

#### **8.16.5.2 Definitions**

The term "DNA shuffling" is used herein to indicate recombination between substantially homologous but non-identical sequences, in some embodiments DNA shuffling may involve crossover via non-homologous recombination, such as via *cer/lox* and/or *flp/frt* systems and the like.

The term "amplification" means that the number of copies of a polynucleotide is increased.

The term "identical" or "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a polynucleotide or the overall polynucleotide are identical or complementary to areas of another polynucleotide or the polynucleotide.

The term "corresponds to" is used herein to mean that a polynucleotide sequence is homologous (*i.e.*, is identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the complementary sequence is homologous to all or a portion of a reference polynucleotide sequence. For illustration, the nucleotide sequence "TATAC" corresponds to a reference "TATAC" and is complementary to a reference sequence "GTATA."

Genetic instability, as used herein, refers to the natural tendency of highly repetitive sequences to be lost through a process of reductive events generally involving

sequence simplification through the loss of repeated sequences. Deletions tend to involve the loss of one copy of a repeat and everything between the repeats.

Quasi-repeated units, as used herein, refers to the repeats to be re-assorted and are by definition not identical. Indeed the method is proposed not only for practically identical encoding units produced by mutagenesis of the identical starting sequence, but also the reassortment of similar or related sequences which may diverge significantly in some regions. Nevertheless, if the sequences contain sufficient homologies to be reassorted by this approach, they can be referred to as "quasi-repeated" units.

Reductive reassortment, as used herein, refers to the increase in molecular diversity that is accrued through deletion (and/or insertion) events that are mediated by repeated sequences.

Repetitive Index (RI), as used herein, is the average number of copies of the quasi-repeated units contained in the cloning vector.

The term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

The term "population" as used herein means a collection of components such as polynucleotides, portions or polynucleotides or proteins. A "mixed population: means a collection of components which belong to the same family of nucleic acids or proteins (*i.e.*, are related) but which differ in their sequence (*i.e.*, are not identical) and hence in their biological activity.

The term "specific polynucleotide" means a polynucleotide having certain end points and having a certain nucleic acid sequence. Two polynucleotides wherein one polynucleotide has the identical sequence as a portion of the second polynucleotide but different ends comprises two different specific polynucleotides.

The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence," "comparison window," "sequence identity," "percentage of sequence identity," and "substantial identity." A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference

sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (*i.e.*, a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

A "comparison window," as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (*i.e.*, gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith and Waterman (1981) Adv. Appl. Math. 2: 482 by the homology alignment algorithm of Needleman and Wuncsch J. Mol. Biol. 48: 443 (1970), by the search of similarity method of Pearson and Lipman Proc. Natl. Acad. Sci. (U.S.A.) 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection, and the best alignment (*i.e.*, resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected.

The term "sequence identity" means that two polynucleotide sequences are identical (*i.e.*, on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (*e.g.*, A, T, C, G, U, or I) occurs in both sequences to yield

the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (*i.e.*, the window size), and multiplying the result by 100 to yield the percentage of sequence identity. This "substantial identity", as used herein, denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence having at least 80 percent sequence identity, preferably at least 85 percent identity, often 90 to 95 percent sequence identity, and most commonly at least 99 percent sequence identity as compared to a reference sequence of a comparison window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

"Conservative amino acid substitutions" refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are : valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

The term "homologous" or "homeologous" means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentrations as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp.

The term "heterologous" means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that areas of polynucleotides or polynucleotides have

areas or regions within their sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are, for example areas of mutations.

The term "cognate" as used herein refers to a gene sequence that is evolutionarily and functionally related between species. For example but not limitation, in the human genome the human CD4 gene is the cognate gene to the mouse 3d4 gene, since the sequences and structures of these two genes indicate that they are highly homologous and both genes encode a protein which functions in signaling T cell activation through MHC class II-restricted antigen recognition.

The term "wild-type" means that the polynucleotide does not comprise any mutations. A "wild type" protein means that the protein will be active at a level of activity found in nature and will comprise the amino acid sequence found in nature.

The term "mutations" means changes in the sequence of a wild-type nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications.

In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

The term "naturally-occurring" as used herein as applied to the object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a

source in nature and which has not been intentionally modified by man in the laboratory is naturally occurring. Generally, the term naturally occurring refers to an object as present in a non-pathological (un-diseased) individual, such as would be typical for the species.

The term "agent" is used herein to denote a chemical compound, a mixture of chemical compounds, an array of spatially localized compounds (*e.g.*, a VLSIPS peptide array, polynucleotide array, and/or combinatorial small molecule array), biological macromolecule, a bacteriophage peptide display library, a bacteriophage antibody (*e.g.*, scFv) display library, a polysome peptide display library, or an extract made from biological materials such as bacteria, plants, fungi, or animal (particular mammalian) cells or tissues. Agents are evaluated for potential activity as anti-neoplastics, anti-inflammatories or apoptosis modulators by inclusion in screening assays described hereinbelow. Agents are evaluated for potential activity as specific protein interaction inhibitors (*i.e.*, an agent which selectively inhibits a binding interaction between two predetermined polypeptides but which does not substantially interfere with cell viability) by inclusion in screening assays described hereinbelow.

As used herein, "substantially pure" means an object species is the predominant species present (*i.e.*, on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

As used herein the term "physiological conditions" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under

typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45 °C and 0.001-10 mM divalent cation (*e.g.*,  $Mg^{++}$ ,  $Ca^{++}$ ); preferably about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (*e.g.*, BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v). Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or non-ionic detergents and/or membrane fractions and/or anti-foam agents and/or scintillants.

"Specific hybridization" is defined herein as the formation of hybrids between a first polynucleotide and a second polynucleotide (*e.g.*, a polynucleotide having a distinct but substantially identical sequence to the first polynucleotide), wherein substantially unrelated polynucleotide sequences do not form hybrids in the mixture.

As used herein, the term "single-chain antibody" refers to a polypeptide comprising a  $V_H$  domain and a  $V_L$  domain in polypeptide linkage, generally linked via a spacer peptide (*e.g.*,  $[Gly-Gly-Gly-Gly-Ser]_x$ ), and which may comprise additional amino acid sequences at the amino- and/or carboxy- termini. For example, a single-chain antibody may comprise a tether segment for linking to the encoding polynucleotide. As an example, a scFv is a single-chain antibody. Single-chain antibodies are generally proteins consisting of one or more polypeptide segments of at least 10 contiguous amino substantially encoded by genes of the immunoglobulin superfamily (*e.g.*, see The Immunoglobulin Gene Superfamily, A.F. Williams and A.N. Barclay, in Immunoglobulin Genes, T. Honjo, F.W. Alt, and THE. Rabbits, eds., (1989) Academic press: San Diego, CA, pp. 361-368, which is incorporated herein by reference), most frequently encoded by a rodent, non-human primate, avian, porcine bovine, ovine, goat, or human heavy chain or light chain gene sequence. A functional single-chain antibody generally contains a sufficient portion

of an immunoglobulin superfamily gene product so as to retain the property of binding to a specific target molecule, typically a receptor or antigen (epitope).

As used herein, the term "complementarity-determining region" and "CDR" refer to the art-recognized term as exemplified by the Kabat and Chothia CDR definitions also generally known as supervariable regions or hypervariable loops (Chothia and Leks (1987) *J. Mol. Biol.* 196; 901; Chothia *et al.* (1989) *Nature* 342; 877; E.A. Kabat *et al.*, Sequences of Proteins of Immunological Interest (national Institutes of Health, Bethesda, MD) (1987); and Tramontano *et al.* (1990) *J. Mol. Biol.* 215; 175). Variable region domains typically comprise the amino-terminal approximately 105-115 amino acids of a naturally-occurring immunoglobulin chain (*e.g.*, amino acids 1-110), although variable domains somewhat shorter or longer are also suitable for forming single-chain antibodies.

An immunoglobulin light or heavy chain variable region consists of a "framework" region interrupted by three hypervariable regions, also called CDR's. The extent of the framework region and CDR's have been precisely defined (*see*, "Sequences of Proteins of Immunological Interest," E. Kabat *et al.*, 4th Ed., U.S. Department of Health and human services, Bethesda, MD (1987)). The sequences of the framework regions of different light or heavy chains are relatively conserved within a specie. As used herein, a "human framework region" is a framework region that is substantially identical (about 85 or more, usually 90-95 or more) to the framework region of a naturally occurring human immunoglobulin. the framework region of an antibody, that is the combined framework regions of the constituent light and heavy chains, serves to position and align the CDR's. The CDR's are primarily responsible for binding to an epitope of an antigen.

As used herein, the term "variable segment" refers to a portion of a nascent peptide which comprises a random, pseudorandom, or defined kernal sequence. A variable segment" refers to a portion of a nascent peptide which comprises a random pseudorandom, or defined kernal sequence. A variable segment can comprise both variant and invariant residue positions, and the degree of residue variation at a variant residue position may be limited: both options are selected at the discretion of the practitioner. Typically, variable segments are about 5 to 20 amino acid residues in length (*e.g.*, 8 to 10), although variable segments may be longer and may comprise antibody portions or receptor



proteins, such as an antibody fragment, a nucleic acid binding protein, a receptor protein, and the like.

As used herein, "random peptide sequence" refers to an amino acid sequence composed of two or more amino acid monomers and constructed by a stochastic or random process. A random peptide can include framework or scaffolding motifs, which may comprise invariant sequences.

As used herein "random peptide library" refers to a set of polynucleotide sequences that encodes a set of random peptides, and to the set of random peptides encoded by those polynucleotide sequences, as well as the fusion proteins contain those random peptides.

As used herein, the term "pseudorandom" refers to a set of sequences that have limited variability, such that, for example, the degree of residue variability at another position, but any pseudorandom position is allowed some degree of residue variation, however circumscribed.

As used herein, the term "defined sequence framework" refers to a set of defined sequences that are selected on a non-random basis, generally on the basis of experimental data or structural data; for example, a defined sequence framework may comprise a set of amino acid sequences that are predicted to form a  $\beta$ -sheet structure or may comprise a leucine zipper heptad repeat motif, a zinc-finger domain, among other variations. A "defined sequence kernel" is a set of sequences which encompass a limited scope of variability. Whereas (1) a completely random 10-mer sequence of the 20 conventional amino acids can be any of  $(20)^{10}$  sequences, and (2) a pseudorandom 10-mer sequence of the 20 conventional amino acids can be any of  $(20)^{10}$  sequences but will exhibit a bias for certain residues at certain positions and/or overall, (3) a defined sequence kernel is a subset of sequences if each residue position was allowed to be any of the allowable 20 conventional amino acids (and/or allowable unconventional amino/imino acids). A defined sequence kernel generally comprises variant and invariant residue positions and/or comprises variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), and the like, either segmentally or over the entire length of the individual selected library member sequence. Defined sequence kernels can refer to

either amino acid sequences or polynucleotide sequences. Of illustration and not limitation, the sequences  $(NNK)_{10}$  and  $(NNM)_{10}$ , wherein N represents A, T, G, or C; K represents G or T; and M represents A or C, are defined sequence kernels.

As used herein "epitope" refers to that portion of an antigen or other macromolecule capable of forming a binding interaction that interacts with the variable region binding body of an antibody. Typically, such binding interaction is manifested as an intermolecular contact with one or more amino acid residues of a CDR.

As used herein, "receptor" refers to a molecule that has an affinity for a given ligand. Receptors can be naturally occurring or synthetic molecules. Receptors can be employed in an unaltered state or as aggregates with other species. Receptors can be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of receptors include, but are not limited to, antibodies, including monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells, or other materials), cell membrane receptors, complex carbohydrates and glycoproteins, enzymes, and hormone receptors.

As used herein "ligand" refers to a molecule, such as a random peptide or variable segment sequence, that is recognized by a particular receptor. As one of skill in the art will recognize, a molecule (or macromolecular complex) can be both a receptor and a ligand. In general, the binding partner having a smaller molecular weight is referred to as the ligand and the binding partner having a greater molecular weight is referred to as a receptor.

As used herein, "linker" or "spacer" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide, and serves to place the two molecules in a preferred configuration, *e.g.*, so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

### 8.16.5.3 Methodology

Nucleic acid shuffling is a method for *in vitro* or *in vivo* homologous recombination of pools of shorter or smaller polynucleotides to produce a polynucleotide or polynucleotides. Mixtures of related nucleic acid sequences or polynucleotides are subjected to sexual PCR to provide random polynucleotides, and reassembled to yield a library or mixed population of recombinant hybrid nucleic acid molecules or polynucleotides.

In contrast to cassette mutagenesis, only shuffling and error-prone PCR allow one to mutate a pool of sequences blindly (without sequence information other than primers).

#### 8.16.5.3.1 Advantage of the Mutagenic Shuffling

The advantage of the mutagenic shuffling of this invention over error-prone PCR alone for repeated selection can best be explained with an example from antibody engineering.

#### 8.16.5.3.2 Inverse Chain Reaction

This method differs from error-prone PCR, in that it is an inverse chain reaction. In error-prone PCR, the number of polymerase start sites and the number of molecules grows exponentially. However, the sequence of the polymerase start sites and the sequence of the molecules remains essentially the same. In contrast, in nucleic acid reassembly or shuffling of random polynucleotides the number of start sites and the number (but not size) of the random polynucleotides decreases over time. For polynucleotides derived from whole plasmids the theoretical endpoint is a single, large concatemeric molecule.

Since cross-overs occur at regions of homology, recombination will primarily occur between members of the same sequence family. This discourages combinations of CDRs that are grossly incompatible (*e.g.*, directed against different epitopes of the same antigen). It is contemplated that multiple families of sequences can be shuffled in the

same reaction. Further, shuffling generally conserves the relative order, such that, for example, CDR1 will not be found in the position of CDR2.

Rare shufflants will contain a large number of the best (eg. highest affinity) CDRs and these rare shufflants may be selected based on their superior affinity.

CDRs from a pool of 100 different selected antibody sequences can be permuted in up to 1006 different ways. This large number of permutations cannot be represented in a single library of DNA sequences. Accordingly, it is contemplated that multiple cycles of DNA shuffling and selection may be required depending on the length of the sequence and the sequence diversity desired.

Error-prone PCR, in contrast, keeps all the selected CDRs in the same relative sequence, generating a much smaller mutant cloud.

#### **8.16.5.3.3 The Template Polynucleotide**

The template polynucleotide which may be used in the methods of this invention may be DNA or RNA. It may be of various lengths depending on the size of the gene or shorter or smaller polynucleotide to be recombined or reassembled. Preferably, the template polynucleotide is from 50 bp to 50 kb. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest can be used in the methods of this invention, and in fact have been successfully used.

The template polynucleotide may be obtained by amplification using the PCR reaction (U.S. Patent No. 4,683,202 and 4,683,195) or other amplification or cloning methods. However, the removal of free primers from the PCR products before subjecting them to pooling of the PCR products and sexual PCR may provide more efficient results. Failure to adequately remove the primers from the original pool before sexual PCR can lead to a low frequency of crossover clones.

The template polynucleotide often should be double-stranded. A double-stranded nucleic acid molecule is recommended to ensure that regions of the resulting

single-stranded polynucleotides are complementary to each other and thus can hybridize to form a double-stranded molecule.

It is contemplated that single-stranded or double-stranded nucleic acid polynucleotides having regions of identity to the template polynucleotide and regions of heterology to the template polynucleotide may be added to the template polynucleotide, at this step. It is also contemplated that two different but related polynucleotide templates can be mixed at this step.

The double-stranded polynucleotide template and any added double-or single-stranded polynucleotides are subjected to sexual PCR which includes slowing or halting to provide a mixture of from about 5 bp to 5 kb or more. Preferably the size of the random polynucleotides is from about 10 bp to 1000 bp, more preferably the size of the polynucleotides is from about 20 bp to 500 bp.

#### **8.16.5.3.4 Use of Double-Stranded Nucleic Acid Having Multiple Nicks**

Alternatively, it is also contemplated that double-stranded nucleic acid having multiple nicks may be used in the methods of this invention. A nick is a break in one strand of the double-stranded nucleic acid. The distance between such nicks is preferably 5 bp to 5 kb, more preferably between 10 bp to 1000 bp. This can provide areas of self-priming to produce shorter or smaller polynucleotides to be included with the polynucleotides resulting from random primers, for example.

The concentration of any one specific polynucleotide will not be greater than 1% by weight of the total polynucleotides, more preferably the concentration of any one specific nucleic acid sequence will not be greater than 0.1% by weight of the total nucleic acid.

The number of different specific polynucleotides in the mixture will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

#### 8.16.5.3.5 Increasing the Heterogeneity of the Mixture of Polynucleotides

At this step single-stranded or double-stranded polynucleotides, either synthetic or natural, may be added to the random double-stranded shorter or smaller polynucleotides in order to increase the heterogeneity of the mixture of polynucleotides.

It is also contemplated that populations of double-stranded randomly broken polynucleotides may be mixed or combined at this step with the polynucleotides from the sexual PCR process and optionally subjected to one or more additional sexual PCR cycles.

Where insertion of mutations into the template polynucleotide is desired, single-stranded or double-stranded polynucleotides having a region of identity to the template polynucleotide and a region of heterology to the template polynucleotide may be added in a 20 fold excess by weight as compared to the total nucleic acid, more preferably the single-stranded polynucleotides may be added in a 10 fold excess by weight as compared to the total nucleic acid.

Where a mixture of different but related template polynucleotides is desired, populations of polynucleotides from each of the templates may be combined at a ratio of less than about 1:100, more preferably the ratio is less than about 1:40. For example, a backcross of the wild-type polynucleotide with a population of mutated polynucleotide may be desired to eliminate neutral mutations (*e.g.*, mutations yielding an insubstantial alteration in the phenotypic property being selected for). In such an example, the ratio of randomly provided wild-type polynucleotides which may be added to the randomly provided sexual PCR cycle hybrid polynucleotides is approximately 1:1 to about 100:1, and more preferably from 1:1 to 40:1.

##### 8.16.5.3.5.1 Denaturing and Re-annealing

The mixed population of random polynucleotides are denatured to form single-stranded polynucleotides and then re-annealed. Only those single-stranded polynucleotides having regions of homology with other single-stranded polynucleotides will re-anneal.

The random polynucleotides may be denatured by heating. One skilled in the art could determine the conditions necessary to completely denature the double-stranded nucleic acid. Preferably the temperature is from 80 °C to 100 °C, more preferably the temperature is from 90 °C to 96 °C. other methods which may be used to denature the polynucleotides include pressure (36) and pH.

The polynucleotides may be re-annealed by cooling. Preferably the temperature is from 20 °C to 75 °C, more preferably the temperature is from 40 °C to 65 °C. If a high frequency of crossovers is needed based on an average of only 4 consecutive bases of homology, recombination can be forced by using a low annealing temperature, although the process becomes more difficult. The degree of renaturation which occurs will depend on the degree of homology between the population of single-stranded polynucleotides.

Renaturation can be accelerated by the addition of polyethylene glycol ("PEG") or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%.

#### 8.16.5.3.5.2 Incubation

The annealed polynucleotides are next incubated in the presence of a nucleic acid polymerase and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). The nucleic acid polymerase may be the Klenow fragment, the Taq polymerase or any other DNA polymerase known in the art.

The approach to be used for the assembly depends on the minimum degree of homology that should still yield crossovers. If the areas of identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30 °C. One skilled in the art could vary the temperature of annealing to increase the number of cross-overs achieved.

The polymerase may be added to the random polynucleotides prior to annealing, simultaneously with annealing or after annealing.

The cycle of denaturation, renaturation and incubation in the presence of polymerase is referred to herein as shuffling or reassembly of the nucleic acid. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 50 times, more preferably the sequence is repeated from 10 to 40 times.

#### **8.16.5.3.6 The Resulting Nucleic Acid**

The resulting nucleic acid is a larger double-stranded polynucleotide of from about 50 bp to about 100 kb, preferably the larger polynucleotide is from 500 bp to 50 kb.

This larger polynucleotides may contain a number of copies of a polynucleotide having the same size as the template polynucleotide in tandem. This concatemeric polynucleotide is then denatured into single copies of the template polynucleotide. The result will be a population of polynucleotides of approximately the same size as the template polynucleotide. The population will be a mixed population where single or double-stranded polynucleotides having an area of identity and an area of heterology have been added to the template polynucleotide prior to shuffling.

These polynucleotides are then cloned into the appropriate vector and the ligation mixture used to transform bacteria.

It is contemplated that the single polynucleotides may be obtained from the larger concatemeric polynucleotide by amplification of the single polynucleotide prior to cloning by a variety of methods including PCR (U.S. Patent No. 4,683,195 and 4,683,202), rather than by digestion of the concatemer.

#### **8.16.5.3.7 Vectors Used for Cloning**

The vector used for cloning is not critical provided that it will accept a polynucleotide of the desired size. If expression of the particular polynucleotide is



desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the polynucleotide to allow expression of the polynucleotide in the host cell. Preferred vectors include the pUC series and the pBR series of plasmids.

#### **8.16.5.3.8 The Resulting Bacterial Population**

The resulting bacterial population will include a number of recombinant polynucleotides having random mutations. This mixed population may be tested to identify the desired recombinant polynucleotides. The method of selection will depend on the polynucleotide desired.

For example, if a polynucleotide which encodes a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the portions of the polynucleotides in the population or library may be tested for their ability to bind to the ligand by methods known in the art (*i.e.* panning, affinity chromatography). If a polynucleotide which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the polynucleotides in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify polynucleotides which confer the desired properties onto the protein.

It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee WI). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

#### **8.16.5.3.9 Cycles of Nucleic Acid Shuffling**

It is further contemplated that a number of cycles of nucleic acid shuffling may be conducted with polynucleotides from a sub-population of the first population, which sub-population contains DNA encoding the desired recombinant protein. In this manner, proteins with even higher binding affinities or enzymatic activity could be achieved.

It is also contemplated that a number of cycles of nucleic acid shuffling may be conducted with a mixture of wild-type polynucleotides and a sub-population of nucleic acid from the first or subsequent rounds of nucleic acid shuffling in order to remove any silent mutations from the sub-population.

#### **8.16.5.3.10 The Starting Nucleic Acid**

Any source of nucleic acid, in purified form can be utilized as the starting nucleic acid. Thus the process may employ DNA or RNA including messenger RNA, which DNA or RNA may be single or double stranded. In addition, a DNA-RNA hybrid which contains one strand of each may be utilized. The nucleic acid sequence may be of various lengths depending on the size of the nucleic acid sequence to be mutated. Preferably the specific nucleic acid sequence is from 50 to 50000 base pairs. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest may be used in the methods of this invention.

The nucleic acid may be obtained from any source, for example, from plasmids such as pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) (U.S. Patent no. 4,683,202 and 4,683,195). Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Any specific nucleic acid sequence can be used to produce the population of hybrids by the present process. It is only necessary that a small population of hybrid

sequences of the specific nucleic acid sequence exist or be created prior to the present process.

#### 8.16.5.3.11 Creation of the Initial Population of Sequences

The initial small population of the specific nucleic acid sequences having mutations may be created by a number of different methods. Mutations may be created by error-prone PCR. Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. Alternatively, mutations can be introduced into the template polynucleotide by oligonucleotide-directed mutagenesis. In oligonucleotide-directed mutagenesis, a short sequence of the polynucleotide is removed from the polynucleotide using restriction enzyme digestion and is replaced with a synthetic polynucleotide in which various bases have been altered from the original sequence. The polynucleotide sequence can also be altered by chemical mutagenesis. Chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, hydrazine or formic acid. Other agents which are analogues of nucleotide precursors include nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. Generally, these agents are added to the PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used. Random mutagenesis of the polynucleotide sequence can also be achieved by irradiation with X-rays or ultraviolet light. Generally, plasmid polynucleotides so mutagenized are introduced into *E. coli* and propagated as a pool or library of hybrid plasmids.

Alternatively the small mixed population of specific nucleic acids may be found in nature in that they may consist of different alleles of the same gene or the same gene from different related species (*i.e.*, cognate genes). Alternatively, they may be related DNA sequences found within one species, for example, the immunoglobulin genes.

Once the mixed population of the specific nucleic acid sequences is generated, the polynucleotides can be used directly or inserted into an appropriate cloning vector, using techniques well-known in the art.

#### **8.16.5.3.11.1 The Choice of Vector**

The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates of this invention may be plasmids, phages, cosmids, phagemids, viruses (*e.g.*, retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (*e.g.*, coat protein, spike glycoprotein, capsid protein). For example, cosmids and phagemids are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large polynucleotides.

#### **8.16.5.3.11.2 Clonal Amplification**

If the mixed population of the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of hybrids does not increase. Utility can be readily determined by screening expressed polypeptides.

#### **8.16.5.3.12 Incorporation of Any Sequence Mixture at Any Specific Position**

The DNA shuffling method of this invention can be performed blindly on a pool of unknown sequences. By adding to the reassembly mixture oligonucleotides (with ends that are homologous to the sequences being reassembled) any sequence mixture can be incorporated at any specific position into another sequence mixture. Thus, it is contemplated that mixtures of synthetic oligonucleotides, PCR polynucleotides or even whole genes can be mixed into another sequence library at defined positions. The insertion of one sequence (mixture) is independent from the insertion of a sequence in another part of the template. Thus, the degree of recombination, the homology required, and the diversity of the library can be independently and simultaneously varied along the length of the reassembled DNA.

This approach of mixing two genes may be useful for the humanization of antibodies from murine hybridomas. The approach of mixing two genes or inserting alternative sequences into genes may be useful for any therapeutically used protein, for example, interleukin I, antibodies, tPA and growth hormone. The approach may also be useful in any nucleic acid for example, promoters or introns or 3' untranslated region or 5' untranslated regions of genes to increase expression or alter specificity of expression of proteins. The approach may also be used to mutate ribozymes or aptamers.

#### 8.16.5.3.13 Creation of Scaffold-like Proteins

Shuffling requires the presence of homologous regions separating regions of diversity. Scaffold-like protein structures may be particularly suitable for shuffling. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta-barrel, and the four-helix bundle which are well-known in the art. This shuffling can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

#### 8.16.5.4 *In vitro* Shuffling

The equivalents of some standard genetic matings may also be performed by shuffling *in vitro*. For example, a "molecular backcross" can be performed by repeatedly mixing the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (*i.e.* immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not, an advantage which cannot be achieved by error-prone mutagenesis or cassette mutagenesis methods.

Large, functional genes can be assembled correctly from a mixture of small random polynucleotides. This reaction may be of use for the reassembly of genes from the highly fragmented DNA of fossils. In addition random nucleic acid fragments from fossils may be combined with polynucleotides from similar genes from related species.

#### 8.16.5.4.1 *In Vitro* Amplification of a Genome

It is also contemplated that the method of this invention can be used for the *in vitro* amplification of a whole genome from a single cell as is needed for a variety of research and diagnostic applications. DNA amplification by PCR is in practice limited to a length of about 40 kb. Amplification of a whole genome such as that of *E. coli* (5, 000 kb) by PCR would require about 250 primers yielding 125 forty kb polynucleotides. This approach is not practical due to the unavailability of sufficient sequence data. On the other hand, random production of polynucleotides of the genome with sexual PCR cycles, followed by gel purification of small polynucleotides will provide a multitude of possible primers. Use of this mix of random small polynucleotides as primers in a PCR reaction alone or with the whole genome as the template should result in an inverse chain reaction with the theoretical endpoint of a single concatemer containing many copies of the genome.

100 fold amplification in the copy number and an average polynucleotide size of greater than 50 kb may be obtained when only random polynucleotides are used. It is thought that the larger concatemer is generated by overlap of many smaller polynucleotides. The quality of specific PCR products obtained using synthetic primers will be indistinguishable from the product obtained from unamplified DNA. It is expected that this approach will be useful for the mapping of genomes.

The polynucleotide to be shuffled can be produced as random or non-random polynucleotides, at the discretion of the practitioner.

#### 8.16.5.5 *In vivo* Shuffling

In an embodiment of *in vivo* shuffling, the mixed population of the specific nucleic acid sequence is introduced into bacterial or eukaryotic cells under conditions such that at least two different nucleic acid sequences are present in each host cell. The polynucleotides can be introduced into the host cells by a variety of different methods. The host cells can be transformed with the smaller polynucleotides using methods known in the art, for example treatment with calcium chloride. If the polynucleotides are inserted into a phage genome, the host cell can be transfected with the recombinant phage genome having the specific nucleic acid sequences. Alternatively, the nucleic acid sequences can be introduced into the host cell using electroporation, transfection, lipofection, biolistics, conjugation, and the like.

In general, in this embodiment, the specific nucleic acids sequences will be present in vectors which are capable of stably replicating the sequence in the host cell. In addition, it is contemplated that the vectors will encode a marker gene such that host cells having the vector can be selected. This ensures that the mutated specific nucleic acid sequence can be recovered after introduction into the host cell. However, it is contemplated that the entire mixed population of the specific nucleic acid sequences need not be present on a vector sequence. Rather only a sufficient number of sequences need be cloned into vectors to ensure that after introduction of the polynucleotides into the host cells each host cell contains one vector having at least one specific nucleic acid sequence present therein. It is also contemplated that rather than having a subset of the population of the specific nucleic acids sequences cloned into vectors, this subset may be already stably integrated into the host cell.

##### 8.16.5.5.1 Homologous Recombination

It has been found that when two polynucleotides which have regions of identity are inserted into the host cells homologous recombination occurs between the two polynucleotides. Such recombination between the two mutated specific nucleic acid sequences will result in the production of double or triple hybrids in some situations.

#### **8.16.5.5.2 Increase in the Frequency of Recombination**

It has also been found that the frequency of recombination is increased if some of the mutated specific nucleic acid sequences are present on linear nucleic acid molecules. Therefore, in a preferred embodiment, some of the specific nucleic acid sequences are present on linear polynucleotides.

#### **8.16.5.5.3 Identification of Host Cell Transformants Containing Desired Sequences**

After transformation, the host cell transformants are placed under selection to identify those host cell transformants which contain mutated specific nucleic acid sequences having the qualities desired. For example, if increased resistance to a particular drug is desired then the transformed host cells may be subjected to increased concentrations of the particular drug and those transformants producing mutated proteins able to confer increased drug resistance will be selected. If the enhanced ability of a particular protein to bind to a receptor is desired, then expression of the protein can be induced from the transformants and the resulting protein assayed in a ligand binding assay by methods known in the art to identify that subset of the mutated population which shows enhanced binding to the ligand. Alternatively, the protein can be expressed in another system to ensure proper processing.

Once a subset of the first recombined specific nucleic acid sequences (daughter sequences) having the desired characteristics are identified, they are then subject to a second round of recombination.

#### **8.16.5.5.4 The Second Cycle of Recombination**

In the second cycle of recombination, the recombined specific nucleic acid sequences may be mixed with the original mutated specific nucleic acid sequences (parent sequences) and the cycle repeated as described above. In this way a set of second recombined specific nucleic acids sequences can be identified which have enhanced characteristics or encode for proteins having enhanced properties. This cycle can be repeated a number of times as desired.



It is also contemplated that in the second or subsequent recombination cycle, a backcross can be performed. A molecular backcross can be performed by mixing the desired specific nucleic acid sequences with a large number of the wild-type sequence, such that at least one wild-type nucleic acid sequence and a mutated nucleic acid sequence are present in the same host cell after transformation. Recombination with the wild-type specific nucleic acid sequence will eliminate those neutral mutations that may affect unselected characteristics such as immunogenicity but not the selected characteristics.

#### **8.16.5.5.5 Generation of a Subset of the Specific Nucleic Acid Sequences**

In another embodiment of this invention, it is contemplated that during the first round a subset of the specific nucleic acid sequences can be generated as smaller polynucleotides by slowing or halting their PCR amplification prior to introduction into the host cell. The size of the polynucleotides must be large enough to contain some regions of identity with the other sequences so as to homologously recombine with the other sequences. The size of the polynucleotides will range from 0.03 kb to 100 kb more preferably from 0.2 kb to 10 kb. It is also contemplated that in subsequent rounds, all of the specific nucleic acid sequences other than the sequences selected from the previous round may be utilized to generate PCR polynucleotides prior to introduction into the host cells.

The shorter polynucleotide sequences can be single-stranded or double-stranded. If the sequences were originally single-stranded and have become double-stranded they can be denatured with heat, chemicals or enzymes prior to insertion into the host cell. The reaction conditions suitable for separating the strands of nucleic acid are well known in the art.

The steps of this process can be repeated indefinitely, being limited only by the number of possible hybrids which can be achieved. After a certain number of cycles, all possible hybrids will have been achieved and further cycles are redundant.

In an embodiment the same mutated template nucleic acid is repeatedly recombined and the resulting recombinants selected for the desired characteristic.

#### **8.16.5.5.6 Cloning into a Vector Capable of Replicating in a Bacteria**

Therefore, the initial pool or population of mutated template nucleic acid is cloned into a vector capable of replicating in a bacteria such as *E. coli*. The particular vector is not essential, so long as it is capable of autonomous replication in *E. coli*. In a preferred embodiment, the vector is designed to allow the expression and production of any protein encoded by the mutated specific nucleic acid linked to the vector. It is also preferred that the vector contain a gene encoding for a selectable marker.

The population of vectors containing the pool of mutated nucleic acid sequences is introduced into the *E. coli* host cells. The vector nucleic acid sequences may be introduced by transformation, transfection or infection in the case of phage. The concentration of vectors used to transform the bacteria is such that a number of vectors is introduced into each cell. Once present in the cell, the efficiency of homologous recombination is such that homologous recombination occurs between the various vectors. This results in the generation of hybrids (daughters) having a combination of mutations which differ from the original parent mutated sequences.

The host cells are then clonally replicated and selected for the marker gene present on the vector. Only those cells having a plasmid will grow under the selection.

#### **8.16.5.5.7 Testing for the Presence of Favorable Mutations**

The host cells which contain a vector are then tested for the presence of favorable mutations. Such testing may consist of placing the cells under selective pressure, for example, if the gene to be selected is an improved drug resistance gene. If the vector allows expression of the protein encoded by the mutated nucleic acid sequence, then such selection may include allowing expression of the protein so encoded, isolation of the protein and testing of the protein to determine whether, for example, it binds with increased efficiency to the ligand of interest.

#### **8.16.5.5.8 Isolation of the Desired Nucleic Acid Sequence**

Once a particular daughter mutated nucleic acid sequence has been identified which confers the desired characteristics, the nucleic acid is isolated either already linked to the vector or separated from the vector. This nucleic acid is then mixed with the first or parent population of nucleic acids and the cycle is repeated.

It has been shown that by this method nucleic acid sequences having enhanced desired properties can be selected.

#### **8.16.5.5.9 Addition of Parental Mutated Sequences to the Cells Containing the First Generation of Hybrids**

In an alternate embodiment, the first generation of hybrids are retained in the cells and the parental mutated sequences are added again to the cells. Accordingly, the first cycle of Embodiment I is conducted as described above. However, after the daughter nucleic acid sequences are identified, the host cells containing these sequences are retained.

The parent mutated specific nucleic acid population, either as polynucleotides or cloned into the same vector is introduced into the host cells already containing the daughter nucleic acids. Recombination is allowed to occur in the cells and the next generation of recombinants, or granddaughters are selected by the methods described above.

This cycle can be repeated a number of times until the nucleic acid or peptide having the desired characteristics is obtained. It is contemplated that in subsequent cycles, the population of mutated sequences which are added to the preferred hybrids may come from the parental hybrids or any subsequent generation.

#### 8.16.5.5.10 "Molecular" Backcross to Eliminate Any Neutral Mutations

In an alternative embodiment, the invention provides a method of conducting a "molecular" backcross of the obtained recombinant specific nucleic acid in order to eliminate any neutral mutations. Neutral mutations are those mutations which do not confer onto the nucleic acid or peptide the desired properties. Such mutations may however confer on the nucleic acid or peptide undesirable characteristics. Accordingly, it is desirable to eliminate such neutral mutations. The method of this invention provide a means of doing so.

In this embodiment, after the hybrid nucleic acid, having the desired characteristics, is obtained by the methods of the embodiments, the nucleic acid, the vector having the nucleic acid or the host cell containing the vector and nucleic acid is isolated.

The nucleic acid or vector is then introduced into the host cell with a large excess of the wild-type nucleic acid. The nucleic acid of the hybrid and the nucleic acid of the wild-type sequence are allowed to recombine. The resulting recombinants are placed under the same selection as the hybrid nucleic acid. Only those recombinants which retained the desired characteristics will be selected. Any silent mutations which do not provide the desired characteristics will be lost through recombination with the wild-type DNA. This cycle can be repeated a number of times until all of the silent mutations are eliminated.

Thus the methods of this invention can be used in a molecular backcross to eliminate unnecessary or silent mutations.

#### 8.16.5.6 Utility

The *in vivo* recombination method of this invention can be performed blindly on a pool of unknown hybrids or alleles of a specific polynucleotide or sequence. However, it is not necessary to know the actual DNA or RNA sequence of the specific polynucleotide.

The approach of using recombination within a mixed population of genes can be useful for the generation of any useful proteins, for example, interleukin I, antibodies, tPA and growth hormone. This approach may be used to generate proteins having altered

specificity or activity. The approach may also be useful for the generation of hybrid nucleic acid sequences, for example, promoter regions, introns, exons, enhancer sequences, 3' untranslated regions or 5' untranslated regions of genes. Thus this approach may be used to generate genes having increased rates of expression. This approach may also be useful in the study of repetitive DNA sequences. Finally, this approach may be useful to mutate ribozymes or aptamers.

Scaffold-like regions separating regions of diversity in proteins may be particularly suitable for the methods of this invention. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta barrel, and the four-helix bundle. The methods of this invention can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

The equivalents of some standard genetic matings may also be performed by the methods of this invention. For example, a "molecular" backcross can be performed by repeated mixing of the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (*i.e.* immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not.

#### 8.16.5.7 Peptide Display Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by peptide display methods, wherein an associated polynucleotide encodes a displayed peptide which is screened for a phenotype (*e.g.*, for affinity for a predetermined receptor (ligand)).

An increasingly important aspect of bio-pharmaceutical drug development and molecular biology is the identification of peptide structures, including the primary amino

acid sequences, of peptides or peptidomimetics that interact with biological macromolecules. one method of identifying peptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (*e.g.*, a receptor), involves the screening of a large library of peptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the peptide.

In addition to direct chemical synthesis methods for generating peptide libraries, several recombinant DNA methods also have been reported. One type involves the display of a peptide sequence, antibody, or other protein on the surface of a bacteriophage particle or cell. Generally, in these methods each bacteriophage particle or cell serves as an individual library member displaying a single species of displayed peptide in addition to the natural bacteriophage or cell protein sequences. Each bacteriophage or cell contains the nucleotide sequence information encoding the particular displayed peptide sequence; thus, the displayed peptide sequence can be ascertained by nucleotide sequence determination of an isolated library member.

A well-known peptide display method involves the presentation of a peptide sequence on the surface of a filamentous bacteriophage, typically as a fusion with a bacteriophage coat protein. The bacteriophage library can be incubated with an immobilized, predetermined macromolecule or small molecule (*e.g.*, a receptor) so that bacteriophage particles which present a peptide sequence that binds to the immobilized macromolecule can be differentially partitioned from those that do not present peptide sequences that bind to the predetermined macromolecule. The bacteriophage particles (*i.e.*, library members) which are bound to the immobilized macromolecule are then recovered and replicated to amplify the selected bacteriophage sub-population for a subsequent round of affinity enrichment and phage replication. After several rounds of affinity enrichment and phage replication, the bacteriophage library members that are thus selected are isolated and the nucleotide sequence encoding the displayed peptide sequence is determined, thereby identifying the sequence(s) of peptides that bind to the predetermined macromolecule (*e.g.*, receptor). Such methods are further described in PCT patent publication Nos. 91/17271, 91/18980, and 91/19818 and 93/08278.

The latter PCT publication describes a recombinant DNA method for the display of peptide ligands that involves the production of a library of fusion proteins with each fusion protein composed of a first polypeptide portion, typically comprising a variable sequence, that is available for potential binding to a predetermined macromolecule, and a second polypeptide portion that binds to DNA, such as the DNA vector encoding the individual fusion protein. When transformed host cells are cultured under conditions that allow for expression of the fusion protein, the fusion protein binds to the DNA vector encoding it. Upon lysis of the host cell, the fusion protein/vector DNA complexes can be screened against a predetermined macromolecule in much the same way as bacteriophage particles are screened in the phage-based display system, with the replication and sequencing of the DNA vectors in the selected fusion protein/vector DNA complexes serving as the basis for identification of the selected library peptide sequence(s).

#### 8.16.5.7.1 Hybrid Methods for Generating Libraries of Peptides and Like Polymers

Other systems for generating libraries of peptides and like polymers have aspects of both the recombinant and *in vitro* chemical synthesis methods. In these hybrid methods, cell-free enzymatic machinery is employed to accomplish the *in vitro* synthesis of the library members (*i.e.*, peptides or polynucleotides). In one type of method, RNA molecules with the ability to bind a predetermined protein or a predetermined dye molecule were selected by alternate rounds of selection and PCR amplification (Tuerk and Gold (1990) Science 249: 505; Ellington and Szostak (1990) Nature 346: 818). A similar technique was used to identify DNA sequences which bind a predetermined human transcription factor (Thiesen and Bach (1990) Nucleic Acids Res. 18: 3203; Beaudry and Joyce (1992) Science 257: 635; PCT patent publication Nos. 92/05258 and 92/14843). In a similar fashion, the technique of *in vitro* translation has been used to synthesize proteins of interest and has been proposed as a method for generating large libraries of peptides. These methods which rely upon *in vitro* translation, generally comprising stabilized polysome complexes, are described further in PCT patent publication Nos. 88/08453, 90/05785, 90/07003, 91/02076, 91/05058, and 92/02536. Applicants have described methods in which library members comprise a fusion protein having a first polypeptide

portion with DNA binding activity and a second polypeptide portion having the library member unique peptide sequence; such methods are suitable for use in cell-free *in vitro* selection formats, among others.

#### 8.16.5.7.2 The Displayed Peptide Sequences

The displayed peptide sequences can be of varying lengths, typically from 3-5000 amino acids long or longer, frequently from 5-100 amino acids long, and often from about 8-15 amino acids long. A library can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Portions or all of the displayed peptide sequence(s) can be random, pseudorandom, defined set kernel, fixed, or the like. The present display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

#### 8.16.5.7.3 Sequence Framework Peptide Libraries

The present invention also provides random, pseudorandom, and defined sequence framework peptide libraries and methods for generating and screening those libraries to identify useful compounds (*e.g.*, peptides, including single-chain antibodies) that bind to receptor molecules or epitopes of interest or gene products that modify peptides or RNA in a desired fashion. The random, pseudorandom, and defined sequence framework peptides are produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.



#### 8.16.5.7.4 Selecting for the Desired Peptide Using Affinity Enrichment

A method of affinity enrichment allows a very large library of peptides and single-chain antibodies to be screened and the polynucleotide sequence encoding the desired peptide(s) or single-chain antibodies to be selected. The polynucleotide can then be isolated and shuffled to recombine combinatorially the amino acid sequence of the selected peptide(s) (or predetermined portions thereof) or single-chain antibodies (or just VHI, VLI or CDR portions thereof). Using these methods, one can identify a peptide or single-chain antibody as having a desired binding affinity for a molecule and can exploit the process of shuffling to converge rapidly to a desired high-affinity peptide or scfv. The peptide or antibody can then be synthesized in bulk by conventional means for any suitable use (*e.g.*, as a therapeutic or diagnostic agent).

A significant advantage of the present invention is that no prior information regarding an expected ligand structure is required to isolate peptide ligands or antibodies of interest. The peptide identified can have biological activity, which is meant to include at least specific binding affinity for a selected receptor molecule and, in some instances, will further include the ability to block the binding of other compounds, to stimulate or inhibit metabolic pathways, to act as a signal or messenger, to stimulate or inhibit cellular activity, and the like.

#### 8.16.5.7.5 Shuffling Sequences Selected by Affinity Screening

The present invention also provides a method for shuffling a pool of polynucleotide sequences selected by affinity screening a library of polysomes displaying nascent peptides (including single-chain antibodies) for library members which bind to a predetermined receptor (*e.g.*, a mammalian proteinaceous receptor such as, for example, a peptidergic hormone receptor, a cell surface receptor, an intracellular protein which binds to other protein(s) to form intracellular protein complexes such as hetero-dimers and the like) or epitope (*e.g.*, an immobilized protein, glycoprotein, oligosaccharide, and the like).

Polynucleotide sequences selected in a first selection round (typically by affinity selection for binding to a receptor (*e.g.*, a ligand)) by any of these methods are pooled and

the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral sequences (*i.e.*, having insubstantial functional effect on binding), such as for example by back-crossing with a wild-type or naturally-occurring sequence substantially identical to a selected sequence to produce native-like functional peptides, which may be less immunogenic. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined receptor (ligand).

Prior to or concomitant with the shuffling of selected sequences, the sequences can be mutagenized. In one embodiment, selected library members are cloned in a prokaryotic vector (*e.g.*, plasmid, phagemid, or bacteriophage) wherein a collection of individual colonies (or plaques) representing discrete library members are produced. Individual selected library members can then be manipulated (*e.g.*, by site-directed mutagenesis, cassette mutagenesis, chemical mutagenesis, PCR mutagenesis, and the like) to generate a collection of library members representing a kernel of sequence diversity based on the sequence of the selected library member. The sequence of an individual selected library member or pool can be manipulated to incorporate random mutation, pseudorandom mutation, defined kernel mutation (*i.e.*, comprising variant and invariant residue positions and/or comprising variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), codon-based mutation, and the like, either segmentally or over the entire length of the individual selected library member sequence. The mutagenized selected library members are then shuffled by *in vitro* and/or *in vivo* recombinatorial shuffling as disclosed herein.

#### **8.16.5.7.6 Peptide Libraries Comprising a Plurality of Individual Library Members**

The invention also provides peptide libraries comprising a plurality of individual library members of the invention, wherein (1) each individual library member of said plurality comprises a sequence produced by shuffling of a pool of selected sequences, and (2) each individual library member comprises a variable peptide segment sequence or single-chain antibody segment sequence which is distinct from the variable peptide segment sequences or single-chain antibody sequences of other individual library members in said plurality (although some library members may be present in more than one copy per library due to uneven amplification, stochastic probability, or the like).

#### **8.16.5.7.7 Product-by-Process**

The invention also provides a product-by-process, wherein selected polynucleotide sequences having (or encoding a peptide having) a predetermined binding specificity are formed by the process of: (1) screening a displayed peptide or displayed single-chain antibody library against a predetermined receptor (*e.g.*, ligand) or epitope (*e.g.*, antigen macromolecule) and identifying and/or enriching library members which bind to the predetermined receptor or epitope to produce a pool of selected library members, (2) shuffling by recombination the selected library members (or amplified or cloned copies thereof) which binds the predetermined epitope and has been thereby isolated and/or enriched from the library to generate a shuffled library, and (3) screening the shuffled library against the predetermined receptor (*e.g.*, ligand) or epitope (*e.g.*, antigen macromolecule) and identifying and/or enriching shuffled library members which bind to the predetermined receptor or epitope to produce a pool of selected shuffled library members.

#### **8.16.5.8 Antibody Display and Screening Methods**

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by antibody display methods, wherein an associated polynucleotide encodes a

displayed antibody which is screened for a phenotype (*e.g.*, for affinity for binding a predetermined antigen (ligand)).

Various molecular genetic approaches have been devised to capture the vast immunological repertoire represented by the extremely large number of distinct variable regions which can be present in immunoglobulin chains. The naturally-occurring germ line immunoglobulin heavy chain locus is composed of separate tandem arrays of variable segment genes located upstream of a tandem array of diversity segment genes, which are themselves located upstream of a tandem array of joining (i) region genes, which are located upstream of the constant region genes. During B lymphocyte development, V-D-J rearrangement occurs wherein a heavy chain variable region gene (VH) is formed by rearrangement to form a fused D segment followed by rearrangement with a V segment to form a V-D-J joined product gene which, if productively rearranged, encodes a functional variable region (VH) of a heavy chain. Similarly, light chain loci rearrange one of several V segments with one of several J segments to form a gene encoding the variable region (VL) of a light chain.

#### **8.16.5.8.1 Sequence Diversity**

The vast repertoire of variable regions possible in immunoglobulins derives in part from the numerous combinatorial possibilities of joining V and i segments (and, in the case of heavy chain loci, D segments) during rearrangement in B cell development. Additional sequence diversity in the heavy chain variable regions arises from non-uniform rearrangements of the D segments during V-D-J joining and from N region addition. Further, antigen-selection of specific B cell clones selects for higher affinity variants having non-germline mutations in one or both of the heavy and light chain variable regions; a phenomenon referred to as "affinity maturation" or "affinity sharpening". Typically, these "affinity sharpening" mutations cluster in specific areas of the variable region, most commonly in the complementarity-determining regions (CDRs).

### 8.16.5.8.2 Prokaryotic Expression Systems

In order to overcome many of the limitations in producing and identifying high-affinity immunoglobulins through antigen-stimulated  $\beta$  cell development (*i.e.*, immunization), various prokaryotic expression systems have been developed that can be manipulated to produce combinatorial antibody libraries which may be screened for high-affinity antibodies to specific antigens. Recent advances in the expression of antibodies in *Escherichia coli* and bacteriophage systems (*see*, "Alternative Peptide Display Methods", *infra*) have raised the possibility that virtually any specificity can be obtained by either cloning antibody genes from characterized hybridomas or by de novo selection using antibody gene libraries (*e.g.*, from Ig CDNA).

Combinatorial libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened as bacteriophage plaques or as colonies of lysogens (Huse *et al.* (1989) Science 246: 1275; Caton and Koprowski (1990) Proc. Natl. Acad. Sci. (U.S.A.) 87: 6450; Mullinax *et al.* (1990) Proc. Natl. Acad. Sci. (U.S.A.) 87: 8095; Persson *et al.* (1991) Proc. Natl. Acad. Sci. (U.S.A.) 88: 2432). Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (Kang *et al.* (1991) Proc. Natl. Acad. Sci. (U.S.A.) 88: 4363; Clackson *et al.* (1991) Nature 352: 624; McCafferty *et al.* (1990) Nature 348: 552; Burton *et al.* (1991) Proc. Natl. Acad. Sci. (U.S.A.) 88: 10134; Hoogenboom *et al.* (1991) Nucleic Acids Res. 19: 4133; Chang *et al.* (1991) J. Immunol. 147: 3610; Breitling *et al.* (1991) Gene 104: 147; Marks *et al.* (1991) J. Mol. Biol. 222@: 581; Barbas *et al.* (1992) Proc. Natl. Acad. Sci. (U.S.A.) 89: 4457; Hawkins and Winter (1992) J. Immunol. 22: 867; Marks *et al.* (1992) Biotechnology 10: 779; Marks *et al.* (1992) J. Biol. Chem. 267: 16007; Lowman *et al.* (1991) Biochemistry 30: 10832; Lerner *et al.* (1992) Science 258: 1313, incorporated herein by reference). Typically, a bacteriophage antibody display library is screened with a receptor (*e.g.*, polypeptide, carbohydrate, glycoprotein, nucleic acid) that is immobilized (*e.g.*, by covalent linkage to a chromatography resin to enrich for reactive phage by affinity chromatography) and/or labeled (*e.g.*, to screen plaque or colony lifts).

### 8.16.5.8.3 Single-Chain Fragment Variable Libraries

One particularly advantageous approach has been the use of so-called single-chain fragment variable (scfv) libraries (Marks *et al.* (1992) Biotechnology 10: 779; Winter G and Milstein C (1991) Nature 349: 293; Clackson *et al.* (1991) op. cit.; Marks *et al.* (1991) J. Mol. Biol. 222: 581; Chaudhary *et al.* (1990) Proc. Natl. Acad. Sci. (USA) 87: 1066; Chiswell *et al.* (1992) TIBTECH 10: 80; McCafferty *et al.* (1990) op.cit.; and Huston *et al.* (1988) Proc. Natl. Acad. Sci. (USA) 85: 5879). Various embodiments of scfv libraries displayed on bacteriophage coat proteins have been described.

Beginning in 1988, single-chain analogues of Fv fragments and their fusion proteins have been reliably generated by antibody engineering methods. The first step generally involves obtaining the genes encoding VH and VL domains with desired binding properties; these V genes may be isolated from a specific hybridoma cell line, selected from a combinatorial V-gene library, or made by V gene synthesis. The single-chain Fv is formed by connecting the component V genes with an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)<sub>3</sub> or equivalent linker peptide(s). The linker bridges the C-terminus of the first V region and N-terminus of the second, ordered as either VH-linker-VL or VL-linker-VH'. In principle, the scfv binding site can faithfully replicate both the affinity and specificity of its parent antibody combining site.

Thus, scfv fragments are comprised of VH and VL domains linked into a single polypeptide chain by a flexible linker peptide. After the scfv genes are assembled, they are cloned into a phagemid and expressed at the tip of the M13 phage (or similar filamentous bacteriophage) as fusion proteins with the bacteriophage PIII (gene 3) coat protein. Enriching for phage expressing an antibody of interest is accomplished by panning the recombinant phage displaying a population scfv for binding to a predetermined epitope (*e.g.*, target antigen, receptor).

The linked polynucleotide of a library member provides the basis for replication of the library member after a screening or selection procedure, and also provides the basis for the determination, by nucleotide sequencing, of the identity of the displayed peptide sequence or VH and VL amino acid sequence. The displayed peptide (s) or single-chain

antibody (e. g., scfv) and/or its VH and VL domains or their CDRs can be cloned and expressed in a suitable expression system. often polynucleotides encoding the isolated VH and VL domains will be ligated to polynucleotides encoding constant regions (CH and CL) to form polynucleotides encoding complete antibodies (e.g., chimeric or fully-human), antibody fragments, and the like. Often polynucleotides encoding the isolated CDRs will be grafted into polynucleotides encoding a suitable variable region framework (and optionally constant regions) to form polynucleotides encoding complete antibodies (e.g., humanized or fully-human), antibody fragments, and the like. Antibodies can be used to isolate preparative quantities of the antigen by immunoaffinity chromatography. Various other uses of such antibodies are to diagnose and/or stage disease (e.g., neoplasia) and for therapeutic application to treat disease, such as for example: neoplasia, autoimmune disease, AIDS, cardiovascular disease, infections, and the like.

#### 8.16.5.8.4 Increasing the Combinatorial Diversity of a SCFV Library

Various methods have been reported for increasing the combinatorial diversity of a scfv library to broaden the repertoire of binding species (idiotype spectrum) The use of PCR has permitted the variable regions to be rapidly cloned either from a specific hybridoma source or as a gene library from non-immunized cells, affording combinatorial diversity in the assortment of VH and VL cassettes which can be combined. Furthermore, the VH and VL cassettes can themselves be diversified, such as by random, pseudorandom, or directed mutagenesis. Typically, VH and VL cassettes are diversified in or near the complementarity-determining regions (CDRS), often the third CDR, CDR3. Enzymatic inverse PCR mutagenesis has been shown to be a simple and reliable method for constructing relatively large libraries of scfv site-directed hybrids (Stemmer *et al.* (1993) Biotechniques 14: 256), as has error-prone PCR and chemical mutagenesis (Deng *et al.* (1994) J. Biol. Chem. 269: 953 3). Riechmann *et al.* (1993) Biochemistry 32: 8848 showed semi-rational design of an antibody scfv fragment using site-directed randomization by degenerate oligonucleotide PCR and subsequent phage display of the resultant scfv hybrids. Barbas *et al.* (1992) on.cit. attempted to circumvent the problem of

limited repertoire sizes resulting from using biased variable region sequences by randomizing the sequence in a synthetic CDR region of a human tetanus toxoid-binding Fab.

CDR randomization has the potential to create approximately  $1 \times 10^{20}$  CDRs for the heavy chain CDR3 alone, and a roughly similar number of variants of the heavy chain CDR1 and CDR2, and light chain CDR1-3 variants. Taken individually or together, the combination possibilities of CDR randomization of heavy and/or light chains requires generating a prohibitive number of bacteriophage clones to produce a clone library representing all possible combinations, the vast majority of which will be non-binding. Generation of such large numbers of primary transformants is not feasible with current transformation technology and bacteriophage display systems. For example, Barbas *et al.* (1992) op.cit. only generated  $5 \times 10^7$  transformants, which represents only a tiny fraction of the potential diversity of a library of thoroughly randomized CDRs.

Despite these substantial limitations, bacteriophage display of scfv have already yielded a variety of useful antibodies and antibody fusion proteins. A bispecific single chain antibody has been shown to mediate efficient tumor cell lysis (Gruber *et al.* (1994) J. Immunol. 152: 5368). Intracellular expression of an anti-Rev scfv has been shown to inhibit HIV-1 virus replication *in vitro* (Duan *et al.* (1994) Proc. Natl. Acad. Sci. (USA) 91: 5075), and intracellular expression of an anti-p21<sup>ras</sup> scfv has been shown to inhibit meiotic maturation of *Xenopus* oocytes (Biocca *et al.* (1993) Biochem. Biophys. Res. Commun. 197: 422). Recombinant scfv which can be used to diagnose HIV infection have also been reported, demonstrating the diagnostic utility of scfv (Lilley *et al.* (1994) J. Immunol. Meth. 171: 211). Fusion proteins wherein an scFv is linked to a second polypeptide, such as a toxin or fibrinolytic activator protein, have also been reported (Holvost *et al.* (1992) Eur. J. Biochem. 210: 945; Nicholls *et al.* (1993) J. Biol. Chem. 268: 5302).



#### 8.16.5.8.5 Use of *in vitro* and *in vivo* Shuffling Methods to Recombine CDRs

If it were possible to generate scfv libraries having broader antibody diversity and overcoming many of the limitations of conventional CDR mutagenesis and randomization methods which can cover only a very tiny fraction of the potential sequence combinations, the number and quality of scfv antibodies suitable for therapeutic and diagnostic use could be vastly improved. To address this, the *in vitro* and *in vivo* shuffling methods of the invention are used to recombine CDRs which have been obtained (typically via PCR amplification or cloning) from nucleic acids obtained from selected displayed antibodies. Such displayed antibodies can be displayed on cells, on bacteriophage particles, on polysomes, or any suitable antibody display system wherein the antibody is associated with its encoding nucleic acid(s). In a variation, the CDRs are initially obtained from mRNA (or cDNA) from antibody-producing cells (*e.g.*, plasma cells/splenocytes from an immunized wild-type mouse, a human, or a transgenic mouse capable of making a human antibody as in W092/03918, W093/12227, and W094/25585), including hybridomas derived therefrom.

Polynucleotide sequences selected in a first selection round (typically by affinity selection for displayed antibody binding to an antigen (*e.g.*, a ligand) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination, especially shuffling of CDRs (typically shuffling heavy chain CDRs with other heavy chain CDRs and light chain CDRs with other light chain CDRS) to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are expressed in a selection format as a displayed antibody and subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection until an antibody of the desired binding affinity is obtained. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral antibody framework sequences (*i.e.*, having insubstantial functional effect on antigen binding), such as for example by back-crossing with a human variable region framework to produce human-like sequence antibodies. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined antigen.

#### **8.16.5.8.6 Controlling the Average Binding Affinity of Selected SCFV Library Members**

Alternatively, or in combination with the noted variations, the valency of the target epitope may be varied to control the average binding affinity of selected scfv library members. The target epitope can be bound to a surface or substrate at varying densities, such as by including a competitor epitope, by dilution, or by other method known to those in the art. A high density (valency) of predetermined epitope can be used to enrich for scfv library members which have relatively low affinity, whereas a low density (valency) can preferentially enrich for higher affinity scfv library members.

#### **8.16.5.8.7 Generating Diverse Variable Segments**

For generating diverse variable segments, a collection of synthetic oligonucleotides encoding random, pseudorandom, or a defined sequence kernel set of peptide sequences can be inserted by ligation into a predetermined site (*e.g.*, a CDR). Similarly, the sequence diversity of one or more CDRs of the single-chain antibody cassette(s) can be expanded by mutating the CDR(s) with site-directed mutagenesis, CDR-replacement, and the like. The resultant DNA molecules can be propagated in a host for cloning and amplification prior to shuffling, or can be used directly (*i.e.*, may avoid loss of diversity which may occur upon propagation in a host cell) and the selected library members subsequently shuffled.

Displayed peptide/polynucleotide complexes (library members) which encode a variable segment peptide sequence of interest or a single-chain antibody of interest are selected from the library by an affinity enrichment technique. This is accomplished by means of a immobilized macromolecule or epitope specific for the peptide sequence of interest, such as a receptor, other macromolecule, or other epitope species. Repeating the affinity selection procedure provides an enrichment of library members encoding the desired sequences, which may then be isolated for pooling and shuffling, for sequencing, and/or for further propagation and affinity enrichment.

The library members without the desired specificity are removed by washing. The degree and stringency of washing required will be determined for each peptide sequence or single-chain antibody of interest and the immobilized predetermined macromolecule or epitope. A certain degree of control can be exerted over the binding characteristics of the nascent peptide/DNA complexes recovered by adjusting the conditions of the binding incubation and the subsequent washing. The temperature, pH, ionic strength, divalent cations concentration, and the volume and duration of the washing will select for nascent peptide/DNA complexes within particular ranges of affinity for the immobilized macromolecule. Selection based on slow dissociation rate, which is usually predictive of high affinity, is often the most practical route. This may be done either by continued incubation in the presence of a saturating amount of free predetermined macromolecule, or by increasing the volume, number, and length of the washes. In each case, the rebinding of dissociated nascent peptide/DNA or peptide/RNA complex is prevented, and with increasing time, nascent peptide/DNA or peptide/RNA complexes of higher and higher affinity are recovered.

Additional modifications of the binding and washing procedures may be applied to find peptides with special characteristics. The affinities of some peptides are dependent on ionic strength or cation concentration. This is a useful characteristic for peptides that will be used in affinity purification of various proteins when gentle conditions for removing the protein from the peptides are required.

One variation involves the use of multiple binding targets (multiple epitope species, multiple receptor species), such that a scfv library can be simultaneously screened for a multiplicity of scfv which have different binding specificities. Given that the size of a scfv library often limits the diversity of potential scfv sequences, it is typically desirable to use scfv libraries of as large a size as possible. The time and economic considerations of generating a number of very large polysome scFv-display libraries can become prohibitive. To avoid this substantial problem, multiple predetermined epitope species (receptor species) can be concomitantly screened in a single library, or sequential screening against a number of epitope species can be used. In one variation, multiple target epitope species, each encoded on a separate bead (or subset of beads), can be mixed

and incubated with a polysome-display scfv library under suitable binding conditions. The collection of beads, comprising multiple epitope species, can then be used to isolate, by affinity selection, scfv library members. Generally, subsequent affinity screening rounds can include the same mixture of beads, subsets thereof, or beads containing only one or two individual epitope species. This approach affords efficient screening, and is compatible with laboratory automation, batch processing, and high throughput screening methods.

#### **8.16.5.8.8 Techniques Used to Diversify a Peptide Library or Single-Chain Antibody Library**

A variety of techniques can be used in the present invention to diversify a peptide library or single-chain antibody library, or to diversify, prior to or concomitant with shuffling, around variable segment peptides found in early rounds of panning to have sufficient binding activity to the predetermined macromolecule or epitope. In one approach, the positive selected peptide/polynucleotide complexes (those identified in an early round of affinity enrichment) are sequenced to determine the identity of the active peptides. Oligonucleotides are then synthesized based on these active peptide sequences, employing a low level of all bases incorporated at each step to produce slight variations of the primary oligonucleotide sequences. This mixture of (slightly) degenerate oligonucleotides is then cloned into the variable segment sequences at the appropriate locations. This method produces systematic, controlled variations of the starting peptide sequences, which can then be shuffled. It requires, however, that individual positive nascent peptide/polynucleotide complexes be sequenced before mutagenesis, and thus is useful for expanding the diversity of small numbers of recovered complexes and selecting variants having higher binding affinity and/or higher binding specificity. In a variation, mutagenic PCR amplification of positive selected peptide/polynucleotide complexes (especially of the variable region sequences, the amplification products of which are shuffled *in vitro* and/or *in vivo* and one or more additional rounds of screening is done prior to sequencing. The same general approach can be employed with single-chain antibodies in order to expand the diversity and enhance the binding affinity/specificity,

typically by diversifying CDRs or adjacent framework regions prior to or concomitant with shuffling. If desired, shuffling reactions can be spiked with mutagenic oligonucleotides capable of *in vitro* recombination with the selected library members can be included. Thus, mixtures of synthetic oligonucleotides and PCR produced polynucleotides (synthesized by error-prone or high-fidelity methods) can be added to the *in vitro* shuffling mix and be incorporated into resulting shuffled library members (shufflants).

#### 8.16.5.8.9 Generation of a Library of CDR-Variant Single-Chain Antibodies

The present invention of shuffling enables the generation of a vast library of CDR-variant single-chain antibodies. One way to generate such antibodies is to insert synthetic CDRs into the single-chain antibody and/or CDR randomization prior to or concomitant with shuffling. The sequences of the synthetic CDR cassettes are selected by referring to known sequence data of human CDR and are selected in the discretion of the practitioner according to the following guidelines: synthetic CDRs will have at least 40 percent positional sequence identity to known CDR sequences, and preferably will have at least 50 to 70 percent positional sequence identity to known CDR sequences. For example, a collection of synthetic CDR sequences can be generated by synthesizing a collection of oligonucleotide sequences on the basis of naturally-occurring human CDR sequences listed in Kabat *et al.* (1991) op. cit. ; the pool (s) of synthetic CDR sequences are calculated to encode CDR peptide sequences having at least 40 percent sequence identity to at least one known naturally-occurring human CDR sequence. Alternatively, a collection of naturally-occurring CDR sequences may be compared to generate consensus sequences so that amino acids used at a residue position frequently (*i.e.*, in at least 5 percent of known CDR sequences) are incorporated into the synthetic CDRs at the corresponding position(s). Typically, several (*e.g.*, 3 to about 50) known CDR sequences are compared and observed natural sequence variations between the known CDRs are tabulated, and a collection of oligonucleotides encoding CDR peptide sequences encompassing all or most permutations of the observed natural sequence variations is synthesized. For example but not for limitation, if a collection of human VH CDR

sequences have carboxy-terminal amino acids which are either Tyr, Val, Phe, or Asp, then the pool(s) of synthetic CDR oligonucleotide sequences are designed to allow the carboxy-terminal CDR residue to be any of these amino acids. In some embodiments, residues other than those which naturally-occur at a residue position in the collection of CDR sequences are incorporated: conservative amino acid substitutions are frequently incorporated and up to 5 residue positions may be varied to incorporate non-conservative amino acid substitutions as compared to known naturally-occurring CDR sequences. Such CDR sequences can be used in primary library members (prior to first round screening) and/or can be used to spike *in vitro* shuffling reactions of selected library member sequences. Construction of such pools of defined and/or degenerate sequences will be readily accomplished by those of ordinary skill in the art.

The collection of synthetic CDR sequences comprises at least one member that is not known to be a naturally-occurring CDR sequence. It is within the discretion of the practitioner to include or not include a portion of random or pseudorandom sequence corresponding to N region addition in the heavy chain CDR; the N region sequence ranges from 1 nucleotide to about 4 nucleotides occurring at V-D and D-J junctions. A collection of synthetic heavy chain CDR sequences comprises at least about 100 unique CDR sequences, typically at least about 1,000 unique CDR sequences, preferably at least about 10,000 unique CDR sequences, frequently more than 50,000 unique CDR sequences; however, usually not more than about  $1 \times 10^6$  unique CDR sequences are included in the collection, although occasionally  $1 \times 10^7$  to  $1 \times 10^8$  unique CDR sequences are present, especially if conservative amino acid substitutions are permitted at positions where the conservative amino acid substituent is not present or is rare (*i.e.*, less than 0.1 percent) in that position in naturally-occurring human CDRS. In general, the number of unique CDR sequences included in a library should not exceed the expected number of primary transformants in the library by more than a factor of 10. Such single-chain antibodies generally bind of about at least  $1 \times 10^{-6}$ , preferably with an affinity of about at least  $5 \times 10^{-7}$  ( $10^{-7}$  M), more preferably with an affinity of at least  $1 \times 10^{-8}$  ( $10^{-8}$  M) to  $1 \times 10^{-9}$  ( $10^{-9}$  M) or more, sometimes up to  $1 \times 10^{-10}$  ( $10^{-10}$  M) or more. Frequently, the predetermined antigen is a human protein, such as for example a human cell surface antigen (e. g., CD4, CD8, IL-2 receptor, EGF receptor, PDGF

receptor), other human biological macromolecule (*e.g.*, thrombomodulin, protein C, carbohydrate antigen, sialyl Lewis antigen, Lselectin), or nonhuman disease associated macromolecule (*e.g.*, bacterial LPS, virion capsid protein or envelope glycoprotein) and the like.

#### 8.16.5.8.10 Expression systems

High affinity single-chain antibodies of the desired specificity can be engineered and expressed in a variety of systems. For example, scfv have been produced in plants (Firek *et al.* (1993) Plant Mol. Biol. 23: 861) and can be readily made in prokaryotic systems (Owens RJ and Young RJ (1994) J. Immunol. Meth. 168: 149; Johnson S and Bird RE (1991) Methods Enzymol 203: 88). Furthermore, the single-chain antibodies can be used as a basis for constructing whole antibodies or various fragments thereof (Kettleborough *et al.* (1994) Eur. J. Immunol. 24: 952). The variable region encoding sequence may be isolated (*e.g.*, by PCR amplification or subcloning) and spliced to a sequence encoding a desired human constant region to encode a human sequence antibody more suitable for human therapeutic uses where immunogenicity is preferably minimized. The polynucleotide(s) having the resultant fully human encoding sequence(s) can be expressed in a host cell (*e.g.*, from an expression vector in a mammalian cell) and purified for pharmaceutical formulation.

The DNA expression constructs will typically include an expression control DNA sequence operably linked to the coding sequences, including naturally-associated or heterologous promoter regions. Preferably, the expression control sequences will be eukaryotic promoter systems in vectors capable of transforming or transfecting eukaryotic host cells. Once the vector has been incorporated into the appropriate host, the host is maintained under conditions suitable for high level expression of the nucleotide sequences, and the collection and purification of the mutant "engineered" antibodies.

As stated previously, the DNA sequences will be expressed in hosts after the sequences have been operably linked to an expression control sequence (*i.e.*, positioned to ensure the transcription and translation of the structural gene). These expression vectors are typically replicable in the host organisms either as episomes or as an integral part of the host chromosomal DNA. Commonly, expression vectors will contain selection

markers, *e.g.*, tetracycline or neomycin, to permit detection of those cells transformed with the desired DNA sequences (*see, e.g.*, U.S. Patent 4,704,362, which is incorporated herein by reference).

#### 8.16.5.8.11 Mammalian Tissue Cell Culture

In addition to eukaryotic microorganisms such as yeast, mammalian tissue cell culture may also be used to produce the polypeptides of the present invention (*see*, Winnacker, "From Genes to Clones," VCH Publishers, *N.i., N.Y.* (1987), which is incorporated herein by reference). Eukaryotic cells are actually preferred, because a number of suitable host cell lines capable of secreting intact immunoglobulins have been developed in the art, and include the CHO cell lines, various COS cell lines, HeLa cells, and myeloma cell lines, but preferably transformed Bcells or hybridomas. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter, an enhancer (Queen *et al.* (1986) *Immunol. Rev.* 89: 49), and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites, and transcriptional terminator sequences. Preferred expression control sequences are promoters derived from immunoglobulin genes, cytomegalovirus, SV40, Adenovirus, Bovine Papilloma Virus, and the like.

Eukaryotic DNA transcription can be increased by inserting an enhancer sequence into the vector. Enhancers are *cis*-acting sequences of between 10 to 300 bp that increase transcription by a promoter. Enhancers can effectively increase transcription when either 51 or 31 to the transcription unit. They are also effective if located within an intron or within the coding sequence itself. Typically, viral enhancers are used, including SV40 enhancers, cytomegalovirus enhancers, polyoma enhancers, and adenovirus enhancers. Enhancer sequences from mammalian systems are also commonly used, such as the mouse immunoglobulin heavy chain enhancer.

Mammalian expression vector systems will also typically include a selectable marker gene. Examples of suitable markers include, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance. The first two marker genes prefer the use of mutant cell lines that lack the ability to grow without the addition of thymidine to the growth medium. Transformed cells can then be



identified by their ability to grow on non-supplemented media. Examples of prokaryotic drug resistance genes useful as markers include genes conferring resistance to G418, mycophenolic acid and hygromycin.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, calcium chloride transfection is commonly utilized for prokaryotic cells, whereas calcium phosphate treatment, lipofection, or electroporation may be used for other cellular hosts. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, liposomes, electroporation, and micro-injection (see, generally, Sambrook *et al.*, *supra*).

Once expressed, the antibodies, individual mutated immunoglobulin chains, mutated antibody fragments, and other immunoglobulin polypeptides of the invention can be purified according to standard procedures of the art, including ammonium sulfate precipitation, fraction column chromatography, gel electrophoresis and the like (see, generally, Scopes, R., Protein Purification, Springer-Verlag, N.Y. (1982)). once purified, partially or to homogeneity as desired, the polypeptides may then be used therapeutically or in developing and performing assay procedures, immunofluorescent stainings, and the like (see, generally, Immunological Methods, Vols. I and II, Eds. Lefkovits and Pernis, Academic Press, New York, N.Y. (1979 and 1981)).

The antibodies generated by the method of the present invention can be used for diagnosis and therapy. By way of illustration and not limitation, they can be used to treat cancer, autoimmune diseases, or viral infections. For treatment of cancer, the antibodies will typically bind to an antigen expressed preferentially on cancer cells, such as erbB-2, CEA, CD33, and many other antigens and binding members well known to those skilled in the art.

#### **8.16.5.9 Yeast Two-Hybrid Screening Assays**

Shuffling can also be used to recombinatorially diversify a pool of selected library members obtained by screening a two-hybrid screening system to identify library members

which bind a predetermined polypeptide sequence. The selected library members are pooled and shuffled by *in vitro* and/or *in vivo* recombination. The shuffled pool can then be screened in a yeast two hybrid system to select library members which bind said predetermined polypeptide sequence (e. g., and SH2 domain) or which bind an alternate predetermined polypeptide sequence (e.g., an SH2 domain from another protein species).

An approach to identifying polypeptide sequences which bind to a predetermined polypeptide sequence has been to use a so-called "two-hybrid" system wherein the predetermined polypeptide sequence is present in a fusion protein (Chien *et al.* (1991) Proc. Natl. Acad. Sci. (USA) 88: 9578). This approach identifies protein-protein interactions *in vivo* through reconstitution of a transcriptional activator (Fields S and Song 0 (1989) Nature 340: 245), the yeast Gal4 transcription protein. Typically, the method is based on the properties of the yeast Gal4 protein, which consists of separable domains responsible for DNA-binding and transcriptional activation. Polynucleotides encoding two hybrid proteins, one consisting of the yeast Gal4 DNA-binding domain fused to a polypeptide sequence of a known protein and the other consisting of the Gal4 activation domain fused to a polypeptide sequence of a second protein, are constructed and introduced into a yeast host cell. Intermolecular binding between the two fusion proteins reconstitutes the Gal4 DNA-binding domain with the Gal4 activation domain, which leads to the transcriptional activation of a reporter gene (*e.g.*, *lacZ*, *HIS3*) which is operably linked to a Gal4 binding site. Typically, the two-hybrid method is used to identify novel polypeptide sequences which interact with a known protein (Silver SC and Hunt SW (1993) Mol. Biol. Rep. 17: 155; Durfee *et al.* (1993) Genes Devel. 7: 555; Yang *et al.* (1992) Science 257: 680; Luban *et al.* (1993) Cell 73: 1067; Hardy *et al.* (1992) Genes Devel. 6: 801; Bartel *et al.* (1993) Biotechniques 14: 920; and Vojtek *et al.* (1993) Cell 74: 205). However, variations of the two-hybrid method have been used to identify mutations of a known protein that affect its binding to a second known protein (Li B and Fields S (1993) FASEB J. 7: 957; Lalo *et al.* (1993) Proc. Natl. Acad. Sci. (USA) 90: 5524; Jackson *et al.* (1993) Mol. Cell. Biol. 13: 2899; and Madura *et al.* (1993) J. Biol. Chem. 268: 12046). Two-hybrid systems have also been used to identify interacting structural domains of two known proteins (Bardwell *et al.* (1993) med. Microbial. 8: 1177; Chakrabarty *et al.* (1992) J. Biol. Chem. 267: 17498; Staudinger *et al.* (1993) J. Biol. Chem. 268: 4608; and Milne GT. and Weaver DT (1993) Genes Devel. 7: 1755) or domains responsible for oligomerization of a single protein (Iwabuchi *et al.* (1993) Oncogene 8: 1693; Bogerd *et al.* (1993) J. Virol. 67: 5030). Variations of two-hybrid

systems have been used to study the *in vivo* activity of a proteolytic enzyme (Dasmahapatra *et al.* (1992) Proc. Natl. Acad. Sci. (USA) 89: 4159). Alternatively, an E. coli/BCCP interactive screening system (Germino *et al.* (1993) Proc. Natl. Acad. Sci. (U.S.A.) 90: 933; Guarente L (1993) Proc. Natl. Acad. Sci. (U.S.A.) 90: 1639) can be used to identify interacting protein sequences (*i.e.*, protein sequences which heterodimerize or form higher order heteromultimers). Sequences selected by a two-hybrid system can be pooled and shuffled and introduced into a two-hybrid system for one or more subsequent rounds of screening to identify polypeptide sequences which bind to the hybrid containing the predetermined binding sequence. The sequences thus identified can be compared to identify consensus sequence(s) and consensus sequence kernels.

In general, standard techniques of recombination DNA technology are described in various publications, *e.g.* Sambrook *et al.*, 1989, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory; Ausubel *et al.*, 1987, Current Protocols in Molecular Biology, vols. 1 and 2 and supplements, and Berger and Kimmel, Methods in Enzymology, Volume 152, Guide to Molecular Cloning Techniques (1987), Academic Press, Inc., San Diego, CA, each of which is incorporated herein in their entirety by reference. Polynucleotide modifying enzymes were used according to the manufacturers recommendations. Oligonucleotides were synthesized on an Applied Biosystems Inc. Model 394 DNA synthesizer using ABI chemicals. If desired, PCR amplimers for amplifying a predetermined DNA sequence may be selected at the discretion of the practitioner.

#### 8.16.5.9.1 Formation of Dimers

One microgram samples of template DNA are obtained and treated with U.V. light to cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template DNA sample. Multiple samples are treated with U.V. light for varying periods of

time to obtain template DNA samples with varying numbers of dimers from U.V. exposure.

#### 8.16.5.9.2 Random Priming Kit

A random priming kit which utilizes a non-proofreading polymerase (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

#### 8.16.5.9.3 Generation of a Selected Mutant Polynucleotide Sequence

The present invention is further directed to a method for generating a selected mutant polynucleotide sequence (or a population of selected polynucleotide sequences) typically in the form of amplified and/or cloned polynucleotides, whereby the selected polynucleotide sequences(s) possess at least one desired phenotypic characteristic (*e.g.*, encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, and the like) which can be selected for. One method for identifying hybrid polypeptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (*e.g.*, a receptor), involves the screening of a large library of polypeptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the polypeptide.

#### **8.16.5.9.4 Generating Libraries Suitable for Affinity Interaction Screening or Phenotypic Screening**

In one embodiment, the present invention provides a method for generating libraries of displayed polypeptides or displayed antibodies suitable for affinity interaction screening or phenotypic screening. The method comprises (1) obtaining a first plurality of selected library members comprising a displayed polypeptide or displayed antibody and an associated polynucleotide encoding said displayed polypeptide or displayed antibody, and obtaining said associated polynucleotides or copies thereof wherein said associated polynucleotides comprise a region of substantially identical sequences, optimally introducing mutations into said polynucleotides or copies, (2) pooling the polynucleotides or copies, (3) producing smaller or shorter polynucleotides by interrupting a random or particularized priming and synthesis process or an amplification process, and (4) performing amplification, preferably PCR amplification, and optionally mutagenesis to homologously recombine the newly synthesized polynucleotides.

#### **8.16.5.9.5 Producing Hybrid Polynucleotides Which Express a Useful Hybrid Polypeptide**

It is a particularly preferred object of the invention to provide a process for producing hybrid polynucleotides which express a useful hybrid polypeptide by a series of steps comprising:

- (a) producing polynucleotides by interrupting a polynucleotide amplification or synthesis process with a means for blocking or interrupting the amplification or synthesis process and thus providing a plurality of smaller or shorter polynucleotides due to the replication of the polynucleotide being in various stages of completion;
- (b) adding to the resultant population of single- or double-stranded polynucleotides one or more single- or double-stranded oligonucleotides, wherein said

added oligonucleotides comprise an area of identity in an area of heterology to one or more of the single- or double-stranded polynucleotides of the population;

(c) denaturing the resulting single- or double-stranded oligonucleotides to produce a mixture of single-stranded polynucleotides, optionally separating the shorter or smaller polynucleotides into pools of polynucleotides having various lengths and further optionally subjecting said polynucleotides to a PCR procedure to amplify one or more oligonucleotides comprised by at least one of said polynucleotide pools;

(d) incubating a plurality of said polynucleotides or at least one pool of said polynucleotides with a polymerase under conditions which result in annealing of said single-stranded polynucleotides at regions of identity between the single-stranded polynucleotides and thus forming of a mutagenized double-stranded polynucleotide chain;

(e) optionally repeating steps (c) and (d);

(f) expressing at least one hybrid polypeptide from said polynucleotide chain, or chains; and

(g) screening said at least one hybrid polypeptide for a useful activity.

In a preferred aspect of the invention, the means for blocking or interrupting the amplification or synthesis process is by utilization of uv light, DNA adducts, DNA binding proteins.

In one embodiment of the invention, the DNA adducts, or polynucleotides comprising the DNA adducts, are removed from the polynucleotides or polynucleotide pool, such as by a process including heating the solution comprising the DNA fragments prior to further processing.

Having thus disclosed exemplary embodiments of the present invention, it should be noted by those skilled in the art that the disclosures are exemplary only and that various other alternatives, adaptations and modifications may be made within the scope of the

present invention. Accordingly, the present invention is not limited to the specific embodiments as illustrated herein.

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following examples are to be considered illustrative and thus are not limiting of the remainder of the disclosure in any way whatsoever.



## References

Unless otherwise indicated, all references cited herein (supra and infra) are incorporated by reference in their entirety.

- |   |
|---|
| Alting-Mecs MA and Short JM: Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts. <i>Gene</i> 137:1, 93-100, 1993.   |
| Arkin AP and Youvan DC: An algorithm for protein engineering: simulations of recursive ensemble mutagenesis. <i>Proc Natl Acad Sci USA</i> 89(16):7811-7815, (Aug 15) 1992.   |
| Arnold FH: Protein engineering for unusual environments. <i>Current Opinion in Biotechnology</i> 4(4):450-455, 1993.  |
| Ausubel FM, et al Editors. <i>Current Protocols in Molecular Biology</i> , Vols. 1 and 2 and supplements. (a.k.a. "The Red Book") Greene Publishing Assoc., Brooklyn, NY, ©1987.  |
| Ausubel FM, et al Editors. <i>Current Protocols in Molecular Biology</i> , Vols. 1 and 2 and supplements. (a.k.a. "The Red Book") Greene Publishing Assoc., Brooklyn, NY, ©1989.  |
| Ausubel FM, et al Editors. <i>Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology</i> . Greene Publishing Assoc., Brooklyn, NY, ©1989.  |
| Ausubel FM, et al Editors. <i>Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology</i> , 2 <sup>nd</sup> Edition. Greene Publishing Assoc., Brooklyn, NY, ©1992.   |
| Barbas CF 3d, Bain JD, Hoekstra DM, Lerner RA: Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. <i>Proc Natl Acad Sci USA</i> 89(10):4457-4461, 1992.  |
| Bardwell AJ, Bardwell L, Johnson DK, Friedberg EC: Yeast DNA recombination and repair proteins Rad1 and Rad10 constitute a complex in vivo mediated by localized hydrophobic domains. <i>Mol Microbiol</i> 8(6):1177-1188, 1993.  |
| Barret AJ, et al., eds.: <i>Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology</i> . San Diego: Academic Press, Inc., 1992.  |
| Bartel P, Chien CT, Sternglanz R, Fields S: Elimination of false positives that arise in using the two-hybrid system. <i>Biotechniques</i> 14(6):920-924, 1993.   |
| Beaudry AA and Joyce GF: Directed evolution of an RNA enzyme. <i>Science</i> 257(5070):635-641, 1992.   |
| Berger and Kimmel, <i>Methods in Enzymology</i> , Volume 152, Guide to Molecular Cloning Techniques. Academic Press, Inc., San Diego, CA, ©1987. (Cumulative Subject Index: Volumes 135-139, 141-167, 1990, 272 pp.)  |
| Bevan M: Binary Agrobacterium vectors for plant transformation. <i>Nucleic Acids Research</i> 12(22):8711-21, 1984.   |
| Biocca S, Pierandrei-Amaldi P, Cattaneo A: Intracellular expression of anti-p21ras single chain Fv fragments inhibits meiotic maturation of xenopus oocytes. <i>Biochem Biophys Res Commun</i> 197(2):422-427, 1993.  |
| Bird et al. <i>Plant Mol Biol</i> 11:651, 1988..  |
| Bogerd HP, Fridell RA, Blair WS, Cullen BR: Genetic evidence that the Tat proteins of human immunodeficiency virus types 1 and 2 can multimerize in the eukaryotic cell nucleus. <i>J Virol</i> 67(8):5030-5034, 1993.  |
| Boyce COL, ed.: <i>Novo's Handbook of Practical Biotechnology</i> . 2 <sup>nd</sup> ed. Bagsvaerd, Denmark, 1986.   |
| Brederode FT, Koper-Zawrtthoff EC, Bol JF: Complete nucleotide sequence of alfalfa mosaic virus RNA 4. <i>Nucleic Acids Research</i> 8(10):2213-23, 1980.   |
| Breitling F, Dubel S, Seehaus T, Klewinghaus I, Little M: A surface expression vector for antibody screening. <i>Gene</i> 104(2):147-153, 1991.   |
| Brown NL, Smith M: Cleavage specificity of the restriction endonuclease isolated from Haemophilus gallinarum (Hga I). <i>Proc Natl Acad Sci U S A</i> 74(8):3213-6, (Aug) 1977.   |
| Burton DR, Barbas CF 3d, Persson MA, Koenig S, Chanock RM, Lerner RA: A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. <i>Proc Natl Acad Sci U S A</i> 88(22):10134-7, (Nov 15) 1991. |
| Caldwell RC and Joyce GF: Randomization of genes by PCR mutagenesis. <i>PCR Methods Appl</i> 2(10):28-33, 1992.   |
| Caton AJ and Koprowski H: Influenze virus hemagglutinin-specific antibodies isolatedf froma   |

- combinatorial expression library are closely related to the immune response of the donor. *Proc Natl Acad Sci USA* 87(16):6450-6454, 1990.
- Chakraborty T, Martin JF, Olson EN: Analysis of the oligomerization of myogenin and E2A products in vivo using a two-hybrid assay system. *J Biol Chem* 267(25):17498-501, 1992.
- Chang CN, Landolfi NF, Queen C: Expression of antibody Fab domains on bacteriophage surfaces. Potential use for antibody selection. *J Immunol* 147(10):3610-4, (Nov 15) 1991.
- Chaudhary VK, Batra JK, Gallo MG, Willingham MC, FitzGerald DJ, Pastan I: A rapid method of cloning functional variable-region antibody genes in *Escherichia coli* as single-chain immunotoxins. *Proc Natl Acad Sci USA* 87(3):1066-1070, 1990.
- Chien CT, Bartel PL, Sternglanz R, Fields S: The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA* 88(21):9578-9582, 1991.
- Chiswell DJ, McCafferty J: Phage antibodies: will new 'coliclonal' antibodies replace monoclonal antibodies? *Trends Biotechnol* 10(3):80-84, 1992.
- Chothia C and Lesk AM: Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196(4):901-917, 1987.
- Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al: Conformations of immunoglobulin hypervariable regions. *Nature* 342(6252):877-883, 1989.
- Clackson T, Hoogenboom HR, Griffiths AD, Winter G: Making antibody fragments using phage display libraries. *Nature* 352(6336):624-628, 1991.
- Conrad M, Topal MD: DNA and spermidine provide a switch mechanism to regulate the activity of restriction enzyme *Nae I*. *Proc Natl Acad Sci USA* 86(24):9707-11, (Dec) 1989.
- Coruzzi G, Broglie R, Edwards C, Chua NH: Tissue-specific and light-regulated expression of a pea nuclear gene encoding the small subunit of ribulose-1,5-bisphosphate carboxylase. *EMBO J* 3(8):1671-9, 1984.
- Dasmahapatra B, DiDomenico B, Dwyer S, Ma J, Sadowski I, Schwartz J: A genetic system for studying the activity of a proteolytic enzyme. *Proc Natl Acad Sci USA* 89(9):4159-4162, 1992.
- Davis LG, Digner MD, Battey JF. Basic Methods in Molecular Biology. Elsevier, New York, NY, ©1986.
- Delegrave S and Youvan DC. *Biotechnology Research* 11:1548-1552, 1993.
- DeLong EF, Wu KY, Prezelin BB, Jovine RV: High abundance of Archaea in Antarctic marine picoplankton. *Nature* 371(6499):695-697, 1994.
- Deng SJ, MacKenzie CR, Sadowska J, Michniewicz J, Young NM, Bundle Dr, Narang SA: Selection of antibody single-chain variable fragments with improved carbohydrate binding by phage display. *J Biol Chem* 269(13):9533-9538, 1994.
- Drauz K, Waldman H, eds.: Enzyme Catalysis in Organic Synthesis: A Comprehensive Handbook. Vol. 1. New York: VCH Publishers, 1995.
- Drauz K, Waldman H, eds.: Enzyme Catalysis in Organic Synthesis: A Comprehensive Handbook. Vol. 2. New York: VCH Publishers, 1995.
- Duan L, Bagasra O, Laughlin MA, Oakes JW, Pomerantz RJ: Potent inhibition of human immunodeficiency virus type 1 replication by an intracellular anti-Rev single-chain antibody. *Proc Natl Acad Sci USA* 91(11):5075-5079, 1994.
- Durfee T, Becherer K, Chen PL, Yeh SH, Yang Y, Kilburn AE, Lee WH, Elledge SJ: The retinoblastoma protein associates with the protein phosphatase type 1 catalytic subunit. *Genes Dev* 7(4):555-569, 1993.
- Ellington AD and Szostak JW: In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818-822, 1990.
- Fields S and Song O: A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245-246, 1989.
- Firek S, Draper J, Owen MR, Gandecha A, Cockburn B, Whitlam GC: Secretion of a functional single-chain Fv protein in transgenic tobacco plants and cell suspension cultures. *Plant Mol Biol* 23(4):861-870, 1993.
- Forsblom S, Rigler R, Ehrenberg M, Philipson L: Kinetic studies on the cleavage of adenovirus DNA by restriction endonuclease *Eco RI*. *Nucleic Acids Res* 3(12):3255-69, (Dec) 1976.
- Foster GD, Taylor SC, eds.: Plant Virology Protocols: From Virus Isolation to Transgenic Resistance. Methods in Molecular Biology, Vol. 81. New Jersey: Humana Press Inc., 1998.
- Franks F, ed.: Protein Biotechnology: Isolation, Characterization, and Stabilization. New Jersey: Humana Press Inc., 1993.

Germineo FJ, Wang ZX, Weissman SM: Screening for in vivo protein-protein interactions. <i>Proc Natl Acad Sci USA</i> 90(3):933-937, 1993.
Gingeras TR, Brooks JE: Cloned restriction/modification system from <i>Pseudomonas aeruginosa</i> . <i>Proc Natl Acad Sci USA</i> 80(2):402-6, 1983 (Jan).
Gluzman Y: SV40-transformed simian cells support the replication of early SV40 mutants. <i>Cell</i> 23(1):175-182, 1981.
Godfrey T, West S, eds.: <i>Industrial Enzymology</i> . 2 <sup>nd</sup> ed. London: Macmillan Press Ltd, 1996.
Gottschalk G: <i>Bacterial Metabolism</i> . 2 <sup>nd</sup> ed. New York: Springer-Verlag Inc., 1986.
Gresshoff PM, ed.: <i>Technology Transfer of Plant Biotechnology</i> . Current Topics in Plant Molecular Biology. Boca Raton: CRC Press, 1997.
Griffin HG, Griffin AM, eds.: <i>PCR Technology: Current Innovations</i> . Boca Raton: CRC Press, Inc., 1994.
Gruber M, Schodin BA, Wilson ER, Kranz DM: Efficient tumor cell lysis mediated by a bispecific single chain antibody expressed in <i>Escherichia coli</i> . <i>J Immunol</i> 152(11):5368-5374, 1994.
Guarente L: Strategies for the identification of interacting proteins. <i>Proc Natl Acad Sci USA</i> 90(5):1639-1641, 1993.
Guilley H, Dudley RK, Jonard G, Balazs E, Richards KE: Transcription of Cauliflower mosaic virus DNA: detection of promoter sequences, and characterization of transcripts. <i>Cell</i> 30(3):763-73, 1982.
Hansen G, Chilton MD: Lessons in gene transfer to plants by a gifted microbe. <i>Curr Top Microbiol Immunol</i> 240:21-57, 1999.
Hardy CF, Sussel L, Shore D: A RAP1-interacting protein involved in transcriptional silencing and telomere length regulation. <i>Genes Dev</i> 6(5):801-814, 1992.
Hartmann HT, et al.: <i>Plant Propagation: Principles and Practices</i> . 6 <sup>th</sup> ed. New Jersey: Prentice Hall, Inc., 1997.
Hawkins RE and Winter G: Cell selection strategies for making antibodies from variable gene libraries: trapping the memory pool. <i>Eur J Immunol</i> 22(3):867-870, 1992.
Holvoet P, Laroche Y, Lijnen HR, Van Hoef B, Brouwers E, De Cock F, Lauwereys M, Gansemans Y, Collen D: Biochemical characterization of single-chain chimeric plasminogen activators consisting of a single-chain Fv fragment of a fibrin-specific antibody and single-chain urokinase. <i>Eur J Biochem</i> 210(3):945-952, 1992.
Honjo T, Alt FW, Rabbitts TH (eds): <i>Immunoglobulin genes</i> . Academic Press: San Diego, CA, pp. 361-368, ©1989.
Hoogenboom HR, Griffiths AD, Johnson KS, Chiswell DJ, Judson P, Winter G: Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. <i>Nucleic Acids Res</i> 19(15):4133-4137, 1991.
Huse WD, Sastry L, Iverson SA, Kang AS, Alting-Mees M, Burton DR, Benkovic SJ, Lerner RA: Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. <i>Science</i> 246(4935):1275-1281, 1989.
Huston JS, Levinson D, Mudgett-Hunter M, Tai MS, Novotney J, Margolies MN, Ridge RJ, Brucoleri RE, Haber E, Crea R, et al: Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in <i>Escherichia coli</i> . <i>Proc Natl Acad Sci USA</i> 85(16):5879-5883, 1988.
Ivan Lefkowitz, Editor. <i>Immunology methods manual : the comprehensive sourcebook of techniques</i> . Academic Press, San Diego, ©1997.
Iwabuchi K, Li B, Bartel P, Fields S: Use of the two-hybrid system to identify the domain of p53 involved in oligomerization. <i>Oncogene</i> 8(6):1693-1696, 1993.
Jackson AL, Pahl PM, Harrison K, Rosamond J, Sclafani RA: Cell cycle regulation of the yeast Cdc7 protein kinase by association with the Dbf4 protein. <i>Mol Cell Biol</i> 13(5):2899-2908, 1993.
Johnson S and Bird RE: <i>Methods Enzymol</i> 203:88, 1991.
Kabat et al: <i>Sequences of Proteins of Immunological Interest</i> , 4th Ed. U.S. Department of Health and Human Services, Bethesda, MD (1987)
Kang AS, Barbas CF, Janda KD, Benkovic SJ, Lerner RA: Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries along phage surfaces. <i>Proc Natl Acad Sci USA</i> 88(10):4363-4366, 1991.
Kettleborough CA, Ansell KH, Allen RW, Rosell-Vives E, Gussow DH, Bendig MM: Isolation of tumor cell-specific single-chain Fv from immunized mice using phage-antibody libraries and the re-construction of whole antibodies from these antibody fragments. <i>Eur J Immunol</i> 24(4):952-958, 1994.

Kruger DH, Barcak GJ, Reuter M, Smith HO: EcoRII can be activated to cleave refractory DNA recognition sites. <i>Nucleic Acids Res</i> 16(9):3997-4008, (May 11) 1988.
Lalo D, Carles C, Sentenac A, Thuriaux P: Interactions between three common subunits of yeast RNA polymerases I and III. <i>Proc Natl Acad Sci USA</i> 90(12):5524-5528, 1993.
Laskowski M Sr: Purification and properties of venom phosphodiesterase. <i>Methods Enzymol</i> 65(1):276-84, 1980.
Lefkovits I and Pernis B, Editors. <i>Immunological Methods</i> , Vols. I and II. Academic Press, New York, NY. Also Vol. III published in Orlando and Vol. IV published in San Diego. ©1979-.
Lerner RA, Kang AS, Bain JD, Burton DR, Barbas CF 3d: Antibodies without immunization. <i>Science</i> 258(5086):1313-1314, 1992.
Leung, D.W., et al, <i>Technique</i> , 1:11-15, 1989.
Li B and Fields S: Identification of mutations in p53 that affect its binding to SV40 large T antigen by using the yeast two-hybrid system. <i>FASEB J</i> 7(10):957-963, 1993.
Lilley GG, Doelzal O, Hillyard CJ, Bernard C, Hudson PJ: Recombinant single-chain antibody peptide conjugates expressed in Escherichia coli for the rapid diagnosis of HIV. <i>J Immunol Methods</i> 171(2):211-226, 1994.
Lowman HB, Bass SH, Simpson N, Wells JA: Selecting high-affinity binding proteins by monovalent phage display. <i>Biochemistry</i> 30(45):10832-10838, 1991.
Luban J, Bossolt KL, Franke EK, Kalpana GV, Goff SP: Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. <i>Cell</i> 73(6):1067-1078, 1993.
Madura K, Dohmen RJ, Varshavsky A: N-recogin/Ubc2 interactions in the N-end rule pathway. <i>J Biol Chem</i> 268(16):12046-54, (Jun 5) 1993.
Marks JD, Griffiths Ad, Malmqvist M, Clackson TP, Bye JM, Winter G: By-passing immunization: building high affinity human antibodies by chain shuffling. <i>Biotechnology (N Y)</i> 10(7):779-783, 1992.
Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G: By-passing immunization. Human antibodies from V-gene libraries displayed on phage. <i>J Mol Biol</i> 222(3):581-597, 1991.
Marks JD, Hoogenboom HR, Griffiths AD, Winter G: Molecular evolution of proteins on filamentous phage. Mimicking the strategy of the immune system. <i>J Biol Chem</i> 267(23):16007-16010, 1992.
Maxam AM, Gilbert W: Sequencing end-labeled DNA with base-specific chemical cleavages. <i>Methods Enzymol</i> 65(1):499-560, 1980.
McCafferty J, Griffiths AD, Winter G, Chiswell DJ: Phage antibodies: filamentous phage displaying antibody variable domains. <i>Nature</i> 348(6301):552-554, 1990.
Method of DNA sequencing.
Miller JH. <u>A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria</u> (see inclusively p. 445). Cold Spring Harbor Laboratory Press, Plainview, NY, ©1992.
Milne GT and Weaver DT: Dominant negative alleles of RAD52 reveal a DNA repair/ recombination complex including Rad51 and Rad52. <i>Genes Dev</i> 7(9):1755-1765, 1993.
Mullinax RL, Gross EA, Amberg JR, Hay BN, Hogrefe HH, Kubitz MM, Greener A, Altling-Mees M, Ardourel D, Short JM, et al: Identification of human antibody fragment clones specific for tetanus toxoid in a bacteriophage lambda immunoexpression library. <i>Proc natl Acad Sci USA</i> 87(20):8095-9099, 1990.
Nath K, Azzolina BA: in <i>Gene Amplification and Analysis</i> (ed. Chirikjian JG), vol. 1, p. 113, Elsevier North Holland, Inc., New York, New York, ©1981.
Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. <i>J Mol Biol</i> 48(3):443-453, 1970.
Nelson M, Christ C, Schildkraut I: Alteration of apparent restriction endonuclease recognition specificities by DNA methylases. <i>Nucleic Acids Res</i> 12(13):5165-73, 1984 (Jul 11).
Nicholls PJ, Johnson VG, Andrew SM, Hoogenboom HR, Raus JC, Youle RJ: Characterization of single-chain antibody (sFv)-toxin fusion proteins produced in vitro in rabbit reticulocyte lysate. <i>J Biol Chem</i> 268(7):5302-5308, 1993.
Oller AR, Vanden Broek W, Conrad M, Topal MD: Ability of DNA and spermidine to affect the activity of restriction endonucleases from several bacterial species. <i>Biochemistry</i> 30(9):2543-9, (Mar 5) 1991.
Owen MRL, Pen J: <u>Transgenic Plants: A Production System for Industrial and Pharmaceutical Proteins</u> . Chichester: John Wiley & Sons, 1996.
Owens RJ and Young RJ: The genetic engineering of monoclonal antibodies. <i>J Immunol Methods</i>

168(2):149-165, 1994.
Pearson WR and Lipman DJ: Improved tools for biological sequence comparison. <i>Proc Natl Acad Sci USA</i> 85(8):2444-2448, 1988.
Pein CD, Reuter M, Meisel A, Cech D, Kruger DH: Activation of restriction endonuclease EcoRII does not depend on the cleavage of stimulator DNA. <i>Nucleic Acids Res</i> 19(19):5139-42, (Oct 11) 1991.
Persson MA, Caothien RH, Burton DR: Generation of diverse high-affinity human monoclonal antibodies by repertoire cloning. <i>Proc Natl Acad Sci USA</i> 88(6):2432-2436, 1991.
Perun TJ, Propst CL, eds.: <i>Computer-Aided Drug Design: Methods and Applications</i> . New York: Marcel Dekker, Inc., 1989.
Qiang BQ, McClelland M, Poddar S, Spokauskas A, Nelson M: The apparent specificity of NotI (5'-GCGGCCGC-3') is enhanced by M.FnuDII or M.BepI methyltransferases (5'-mCGCG-3'): cutting bacterial chromosomes into a few large pieces. <i>Gene</i> 88(1):101-5, (Mar 30) 1990.
Queen C, Foster J, Stauber C, Stafford J: Cell-type specific regulation of a kappa immunoglobulin gene by promoter and enhance elements. <i>Immunol Rev</i> 89:49-68, 1986.
Raleigh EA, Wilson G: Escherichia coli K-12 restricts DNA containing 5-methylcytosine. <i>Proc Natl Acad Sci USA</i> 83(23):9070-4, (Dec) 1986.
Reidhaar-Olson JF and Sauer RT: Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. <i>Science</i> 241(4861):53-57, 1988.
Riechmann L and Weill M: Phage display and selection of a site-directed randomized single-chain antibody Fv fragment for its affinity improvement. <i>Biochemistry</i> 32(34):8848-8855, 1993.
Roberts RJ, Macelis D: REBASE--restriction enzymes and methylases. <i>Nucleic Acids Res</i> 24(1):223-35, (Jan 1) 1996.
Ryan AJ, Royal CL, Hutchinson J, Shaw CH: Genomic sequence of a 12S seed storage protein from oilseed rape ( <i>Brassica napus</i> c.v. jet neuf). <i>Nucl Acids Res</i> 17(9):3584, 1989.
Sambrook J, Fritsch EF, Maniatis T. <i>Molecular Cloning: A Laboratory Manual</i> . Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ©1982.
Sambrook J, Fritsch EF, Maniatis T. <i>Molecular Cloning: A Laboratory Manual</i> . Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ©1989.
Scopes RK. <i>Protein Purification: Principles and Practice</i> . Springer-Verlag, New York, NY, © 1982.
Segel IH: <i>Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems</i> . New York: John Wiley & Sons, Inc., 1993.
Silver SC and Hunt SW 3d: Techniques for cloning cDNAs encoding interactive transcriptional regulatory proteins. <i>Mol Biol Rep</i> 17(3):155-165, 1993.
Smith TF, Waterman MS, Fitch WM: Comparative biosequence metrics. <i>J Mol Evol</i> S18(1):38-46, 1981.
Smith TF, Waterman MS. <i>Adv Appl Math</i> 2: 482-end of article, 1981.
Smith TF, Waterman MS: Identification of common molecular subsequences. <i>J Mol Biol</i> 147(1):195-7, (Mar 25) 1981.
Smith TF, Waterman MS: Overlapping genes and information theory. <i>J Theor Biol</i> 91(2):379-80, (Jul 21) 1981.
Staudinger J, Perry M, Elledge SJ, Olson EN: Interactions among vertebrate helix-loop-helix proteins in yeast using the two-hybrid system. <i>J Biol Chem</i> 268(7):4608-4611, 1993.
Stemmer WP, Morris SK, Wilson BS: Selection of an active single chain Fv antibody from a protein linker library prepared by enzymatic inverse PCR. <i>Biotechniques</i> 14(2):256-265, 1993.
Stemmer WP: DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. <i>Proc Natl Acad Sci USA</i> 91(22):10747-10751, 1994.
Sun D, Hurley LH: Effect of the (+)-CC-1065-(N3-adenine)DNA adduct on in vitro DNA synthesis mediated by Escherichia coli DNA polymerase. <i>Biochemistry</i> 31:10, 2822-9, (Mar 17) 1992,
Tague BW, Dickinson CD, Chrispeels MJ: A short domain of the plant vacuolar protein phytohemagglutinin targets invertase to the yeast vacuole. <i>Plant Cell</i> 2(6):533-46, (June) 1990.
Takahashi N, Kobayashi I: Evidence for the double-strand break repair model of bacteriophage lambda recombination. <i>Proc Natl Acad Sci USA</i> 87(7):2790-4, (Apr) 1990.
Thiesen HJ and Bach C: Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. <i>Nucleic Acids Res</i> 18(11):3203-3209, 1990.
Thomas M, Davis RW: Studies on the cleavage of bacteriophage lambda DNA with EcoRI Restriction endonuclease. <i>J Mol Biol</i> 91(3):315-28, (Jan 25) 1975.

Tingey SV, Walker EL, Corruzzi GM: Glutamine synthetase genes of pea encode distinct polypeptides which are differentially expressed in leaves, roots and nodules. <i>EMBO J</i> 6(1):1-9, 1987.
Topal MD, Thresher RJ, Conrad M, Griffith J: NaeI endonuclease binding to pBR322 DNA induces looping. <i>Biochemistry</i> 30(7):2006-10, (Feb. 19) 1991.
Tramontano A, Chothia C, Lesk AM: Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. <i>J Mol Biol</i> 215(1):175-182, 1990.
Tuerk C and Gold L: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. <i>Science</i> 249(4968):505-510, 1990.
USPN 4,683,195; Filed Feb. 7, 1986, Issued Jul 28, 1987. Mullis KB, Erlich HA, Arnheim N, Horn GT, Saiki RK, Scharf SJ: Process for Amplifying, Detecting, and/or Cloning Nucleic Acid Sequences.
USPN 4,683,202; Filed Oct. 25, 1985, Issued Jul. 28, 1987. Mullis KB: Process for Amplifying Nucleic Acid Sequences.
USPN 4,704,362; Filed Nov. 5, 1979, Issued Nov. 3, 1987. Itakura K, Riggs AD: Recombinant Cloning Vehicle Microbial Polypeptide Expression.
USPN 4,713,337; Filed Jan. 3, 1985, Issued Dec. 15, 1987. Jasin M, Schimmel PR: Method for deletion of a gene from a bacteria.
USPN 4,732,856; Filed April 3, 1984, Issued March 22, 1988. Federoff NV: Transposable elements and process for using same.
USPN 4,963,487; Filed Sept. 14, 1987, Issued Jan. 16, 1990. Schimmel PR: Method for deletion of a gene from a bacteria.
USPN 5,354,656; Filed Oct. 2, 1989, Issued Oct. 11, 1994. Sorge, Joseph A. ; Huse, William D.:
USPN 5,385,835; Filed May 19, 1994, Issued Jan. 31, 1995. Helentjaris, Timothy ; Nienhuis, James: Identification and localization and introgression into plants of desired multigenic traits.
USPN 5,453,247; Filed Nov. 23, 1993, Issued Sept. 26, 1995. Beavis, Ronald C. ; Chait, Brian T.: Instrument and method for the sequencing of genome.
USPN 5,604,100; Filed July 19, 1995, Issued Feb. 18, 1997. Perlin, Mark W.: Method and system for sequencing genomes.
USPN 5,670,321; Filed May 10, 1995, Issued Sept. 23, 1997. Kimmel, Bruce E. ; Ellis, Michael ; Ruddy, David: Efficient method to conduct large-scale genome sequencing.
USPN 5,925,808; Filed Dec. 19, 1997, Issued July 20, 1999. Oliver, Melvin John ; Quisenberry, Jerry Edwin ; Trolinder, Norma Lee Glover ; Keim, Don Lee: Control Of Plant Gene Expression.
USPN 5,953,727; Filed March 6, 1997, Issued Sept. 14, 1999. Maslyn, Timothy J. ; Au-Young, Janice ; Hillman, Jennifer L. ; Hibbert, Harold ; Akerblom, Ingrid E. ; Cheng, Rachel J. ; Tang, Yuanhua T.:Project-based full-length biomolecular sequence database.
USPN 5,965,443; Filed Sept. 9, 1996, Issued Oct. 12, 1999. Reznikoff WS, Goryshin IY: System for in vitro transposition.
USPN 5,981,177; Filed Jan. 25, 1995, Issued Nov. 9, 1999. Demirjian DC, Casadaban MJ, Weber M, Gaines GL: Protein fusion method and constructs.
USPN 5,994,058; Filed March 20, 1995, Issued Nov. 30, 1999. Senapathy, Periannan:Method For Contiguous Genome Sequencing.
USPN 6,023,659; Filed March 6, 1997, Issued Feb. 8, 2000. Seilhamer, Jeffrey J. ; Akerblom, Ingrid E. ; Altus, Christina M. ; Klingler, Tod M. ; Russo, Frank ; Au-Young, Janice ; Hillman, Jennifer L. ; Maslyn, Timothy J.: Database System Employing Protein Function Hierarchies For Viewing Biomolecular Sequence Data.
van de Poll ML, Lafleur MV, van Gog F, Vrieling H, Meerman JH: N-acetylated and deacetylated 4'-fluoro-4-aminobiphenyl and 4-aminobiphenyl adducts differ in their ability to inhibit DNA replication of single-stranded M13 in vitro and of single-stranded phi X174 in Escherichia coli. <i>Carcinogenesis</i> 13(5):751-8, (May) 1992.
Vojtek AB, Hollenberg SM, Cooper JA: Mammalian Ras interacts directly with the serine/threonine kinase Raf. <i>Cell</i> 74(1):205-214, 1993.
Wenzler H, Mignery G, Fisher L, Park W: Sucrose-regulated expression of a chimeric potato tuber gene in leaves of transgenic tobacco plants. <i>Plant Mol Biol</i> 13(4):347-54, 1989.
White JS, White DC: <u>Source Book of Enzymes</u> . Boca Raton: CRC Press, 1997.
Williams and Barclay, in <u>Immunoglobulin Genes, The Immunoglobulin Gene Superfamily</u>
Winnacker EL. <u>From Genes to Clones: Introduction to Gene Technology</u> . VCH Publishers, New York,

NY, ©1987.
Winter G and Milstein C: Man-made antibodies. <i>Nature</i> 349(6307):293-299, 1991.
WO 00/04190; Filed July 15, 1999, Published Jan. 27, 2000. Del Cardayre S, Tobin M, Stemmer WP, Ness JE, Minshull J, Patten PA, Subramanian V, Castle LA, Krebber CM, Bass S, Zhang Y, Cox T, Huisman G, Yuan L, Affholter JA: Evolution of whole cells and organisms by recursive sequence recombination.
WO 00/09755; Filed Aug. 12, 1999, Published Feb. 24, 2000. Zarling D, Reddy G, Pati S: Domain specific gene evolution.
WO 88/08453; Filed Apr. 14, 1988, Published Nov. 3, 1988. Alakhov JB, Baranov, VI, Ovodov SJ, Ryabova LA, Spirin AS: Method of Obtaining Polypeptides in Cell-Free Translation System.
WO 90/05785; Filed Nov. 15, 1989, Published May 31, 1990. Schultz P: Method for Site-Specifically Incorporating Unnatural Amino Acids into Proteins.
WO 90/07003; Filed Jan. 27, 1989, Published June 28, 1990. Baranov VI, Morozov II, Spirin AS: Method for Preparative Expression of Genes in a Cell-free System of Conjugated Transcription/translation.
WO 91/02076; Filed June 14, 1990, Published Feb. 21, 1991. Baranov VI, Ryabova LA, Yarchuk OB, Spirin AS: Method for Obtaining Polypeptides in a Cell-free System.
WO 91/05058; Filed Oct. 5, 1989, Published Apr. 18, 1991. Kawasaki G: Cell-free Synthesis and Isolation of Novel Genes and Polypeptides.
WO 91/17271; Filed May 1, 1990, Published Nov. 14, 1991. Dower WJ, Cwirla SE: Recombinant Library Screening Methods.
WO 91/18980; Filed May 13, 1991, Published Dec. 12, 1991. Devlin JJ: Compositions and Methods for Identifying Biologically Active Molecules.
WO 91/19818; Filed June 20, 1990, Published Dec. 26, 1991. Dower WJ, Cwirla SE, Barrett RW: Peptide Library and Screening Systems.
WO 92/02536; Filed Aug. 1, 1991, Published Feb. 20, 1992. Gold L, Tuerk C: Systematic Polypeptide Evolution by Reverse Translation.
WO 92/03918; Filed Aug. 28, 1991, Published Mar. 19, 1992. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.
WO 92/05258; Filed Sept. 17, 1991, Published Apr. 2, 1992. Fincher GB: Gene Encoding Barley Enzyme.
WO 92/14843; Filed Feb. 21, 1992, Published Sept. 3, 1992. Toole JJ, Griffin LC, Bock LC, Latham JA, Muenchau DD, Krawczyk S: Aptamers Specific for Biomolecules and Method of Making.
WO 93/08278; Filed Oct. 15, 1992, Published Apr. 29, 1993. Schatz PJ, Cull MG, Miller JF, Stemmer WP: Peptide Library and Screening Method.
WO 93/12227; Filed Dec. 17, 1992, Published June 24, 1993. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.
WO 94/25585; Filed Apr. 25, 1994, Published Nov. 10, 1994. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.
WO 95/00530; Filed June 6, 1994, Published Jan. 1, 1995. Fodor, Stephen, P., A.; Lipshutz, Robert, J.; Huang, Xiaohua; Jevons, Luis, Carlos: Hybridization and Sequencing of Nucleic Acids.
WO 96/21031; Filed June 7, 1995, Published July 11, 1996. Tricoli, David, M.; Carney, Kim, J.; Russell, Paul, F.; Quemada, Hector, D.; McMaster, J., Russell; Reynolds, John, F.; Deng, Rosaline, Z.: Transgenic Plants Expressing DNA Constructs Containing A Plurality Of Genes To Impart Virus Resistance.
WO 96/27025; Filed Feb. 21, 1996, Published Sept. 6, 1996. Rabani, Ely, Michael: Device, Compounds, Algorithms, And Methods Of Molecular Characterization And Manipulation With Molecular Parallelism.
WO 97/17429; Filed Nov. 8, 1996, Published May 15, 1997. Oglevee-O'donovan, Wendy; Arteca, Richard, N.; Arteca, Jeannette; Stoots, Eleanor: Method For The Commercial Production Of Transgenic Plants.
WO 97/35966; Filed March 20, 1997, Published Oct. 2, 1997. Minshull J, Stemmer WP: Methods and compositions for cellular and metabolic engineering.
WO 97/37041; Filed March 18, 1997, Published Oct. 9, 1997. Köster, Hubert: DNA Sequencing By Mass Spectrometry.
WO 97/42348; Filed May 5, 1997, Published Nov. 13, 1997. Köster, Hubert; Van Den Boom, Dirk; Ruppert, Andreas: Process For Direct Sequencing During Template Amplification.
WO 98/26407; Filed Dec. 11, 1997, Published June 18, 1998. Sabatini, Cathryn, E.; Heath, Joe, Don; Covitz, Peter, A.; Klinger, Tod, M.; Russo, Frank, D.; Berry, Stephanie, F.: Database And System For Storing, Comparing And Displaying Genomic Information.
WO 98/26408; Filed Dec. 11, 1997, Published June 18, 1998. Sabatini, Cathryn, E.; Heath, Joe, Don; Covitz, Peter, A.; Klinger, Tod, M.; Russo, Frank, D.; Berry, Stephanie, F.: Database And System For

Determining, Storing And Displaying Gene Locus Information.
WO 98/31833; Filed Dec. 12, 1997, Published July 23, 1998. Ju, Jingyue: Nucleic Acid Sequencing With Solid Phase Capturable Terminators.
WO 98/31834; Filed Dec. 12, 1997, Published July 23, 1998. Ju, Jingyue: Sets Of Labeled Energy Transfer Fluorescent Primers And Their Use In Multi Component Analysis.
WO 98/31837; Filed Jan. 16, 1998, Published July 23, 1998. Delcardayre SB, Tobin MB, Stemmer WP, Ness JE, Minshull J, Patten P: Evolution of whole cells and organisms by recursive sequence recombination.
WO 98/36085; Filed Feb. 13, 1998, Published Aug. 20, 1998. Sutliff, Thomas, D. ; Rodriguez, Raymond, L.: Production Of Mature Proteins In Plants.
WO 98/37223; Filed Feb. 18, 1998, Published Aug. 27, 1998. Pang, Sheng-Zhi ; Gonsalves, Dennis ; Jan, Fuh-Jyh: DNA Construct To Confer Multiple Traits On Plants.
WO 99/35494; Filed Jan. 8, 1999, Published July 15, 1999. Tally FP, Tao J, Wendler PA, Connelly G, Gallant PL: Method for identifying validated target and assay combinations.
WO 99/37755; Filed Dec. 11, 1998, Published July 29, 1999. Pati S, Zarling David, Lehman CW, Zeng H: The use of consensus sequences for targeted homologous gene isolation and recombination in gene families.
WO 99/49403; Filed March 25, 1999, Published Sept. 30, 1999. Lincoln, Stephen, E. ; Hodgson, David, M. ; Spiro, Peter, A. ; Russo, Frank, D. ; Akerblom, Ingrid, E. ; Hillman, Jennifer, L. ; Jones, Anissa, Lee ; Bratcher, Shawn, Robert ; Cohen, Howard, Jerome ; Dufour, Gerard ; Wood, Michael, Peter ; Koleszar, Alexander, George ; Banville, Steven, C.: System And Methods For Analyzing Biomolecular Sequences.
WO95/11995; Filed Oct. 26, 1994, Published May 4, 1995. Chee M, Cronin MT, Fodor SP, Gingeras TR, Huang XC, Hubbell EA, Lipshutz RJ, Lobban PE, Miyada CG, Morris MS, Shah N, Sheldon EL: Arrays Of Nucleic Acid Probes On Biological Chips.
Wong CH, Whitesides GM: <u>Enzymes in Synthetic Organic Chemistry</u> . Vol. 12. New York: Elsevier Science Publications, 1995.
Yang X, Hubbard EJ, Carlson M: A protein kinase substrate identified by the two-hybrid system. <i>Science</i> 257(5070):680-2, (Jul 31) 1992.

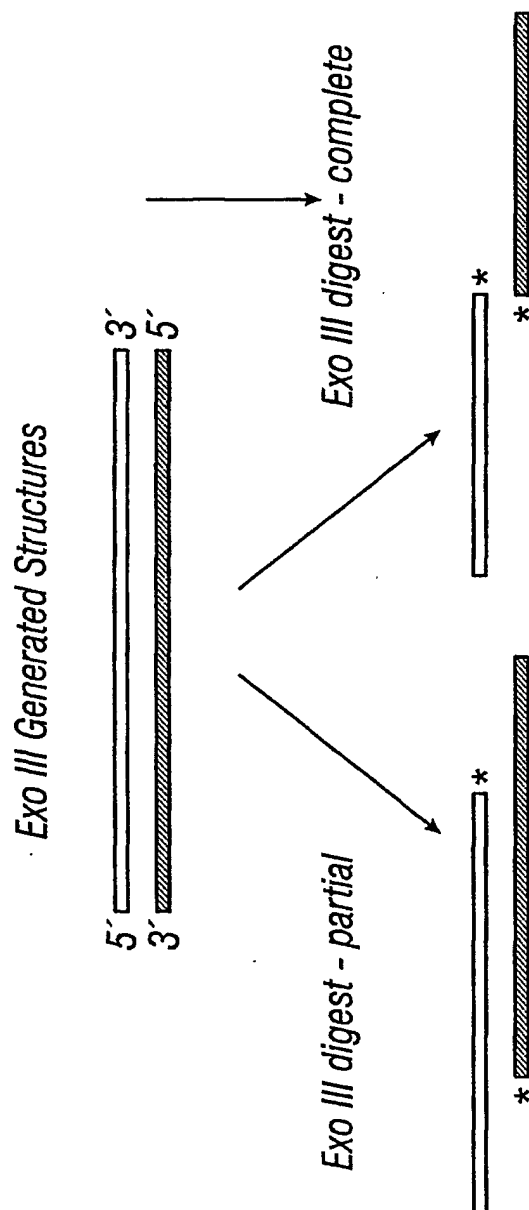


What is claimed is:

1. A method of producing an improved organism having a desirable trait comprising:  
a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence of said improved organism.
2. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 15 different genes.
3. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 50 different genes.
4. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 100 different genes.
5. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 15 different genes.
6. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 50 different genes.
7. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 100 different genes.
8. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an alteration in the expression of at least 15 different genes.
9. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an alteration in the expression of at least 50 different genes.

10. The method of claim 1, wherein the set of substantial genetic mutations in step b) is comprised of an alteration in the expression of at least 100 different genes.
11. A method of producing an improved organism having a desirable trait comprising:  
a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms each having at least one genetic mutation, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations c) detecting the manifestation of at least two genetic mutations, d) introducing at least two detected genetic mutations into one organism, and e) optionally repeating any of steps a), b), c), and d).
12. The method of claim 11, wherein step d) is comprised of a knocking out of at least 15 different genes in one organism.
13. The method of claim 11, wherein step d) is comprised of a knocking out of at least 50 different genes in one organism.
14. The method of claim 11, wherein step d) is comprised of a knocking out of at least 100 different genes in one organism.
15. The method of claim 11, wherein step d) is comprised of an introduction of at least 15 different genes into one organism.
16. The method of claim 11, wherein step d) is comprised of an introduction of at least 50 different genes into one organism.
17. The method of claim 11, wherein step d) is comprised of an introduction of at least 100 different genes into one organism.

18. The method of claim 11, wherein step d) is comprised of an alteration in the expression of at least 15 different genes in one organism.
19. The method of claim 11, wherein step d) is comprised of an alteration in the expression of at least 50 different genes in one organism.
20. The method of claim 11, wherein step d) is comprised of an alteration in the expression of at least 100 different genes in one organism.
21. A method for identifying a gene that alters a trait of an organism, comprising: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence an organism having said altered trait, and d) determining the nucleotide sequence of a gene that has been mutagenized in the organism having the altered trait.
22. A method for producing an organism with an improved trait, comprising: a) functionally knocking out an endogenous gene in a substantially clonal population of organisms; b) transferring a library of altered genes into the substantially clonal population of organisms, wherein each altered gene differs from the endogenous gene at only one codon; c) detecting a mutagenized organism having an improved trait; and d) determining the nucleotide sequence of an gene that has been transferred into the detected organism.



**FIG. 1**

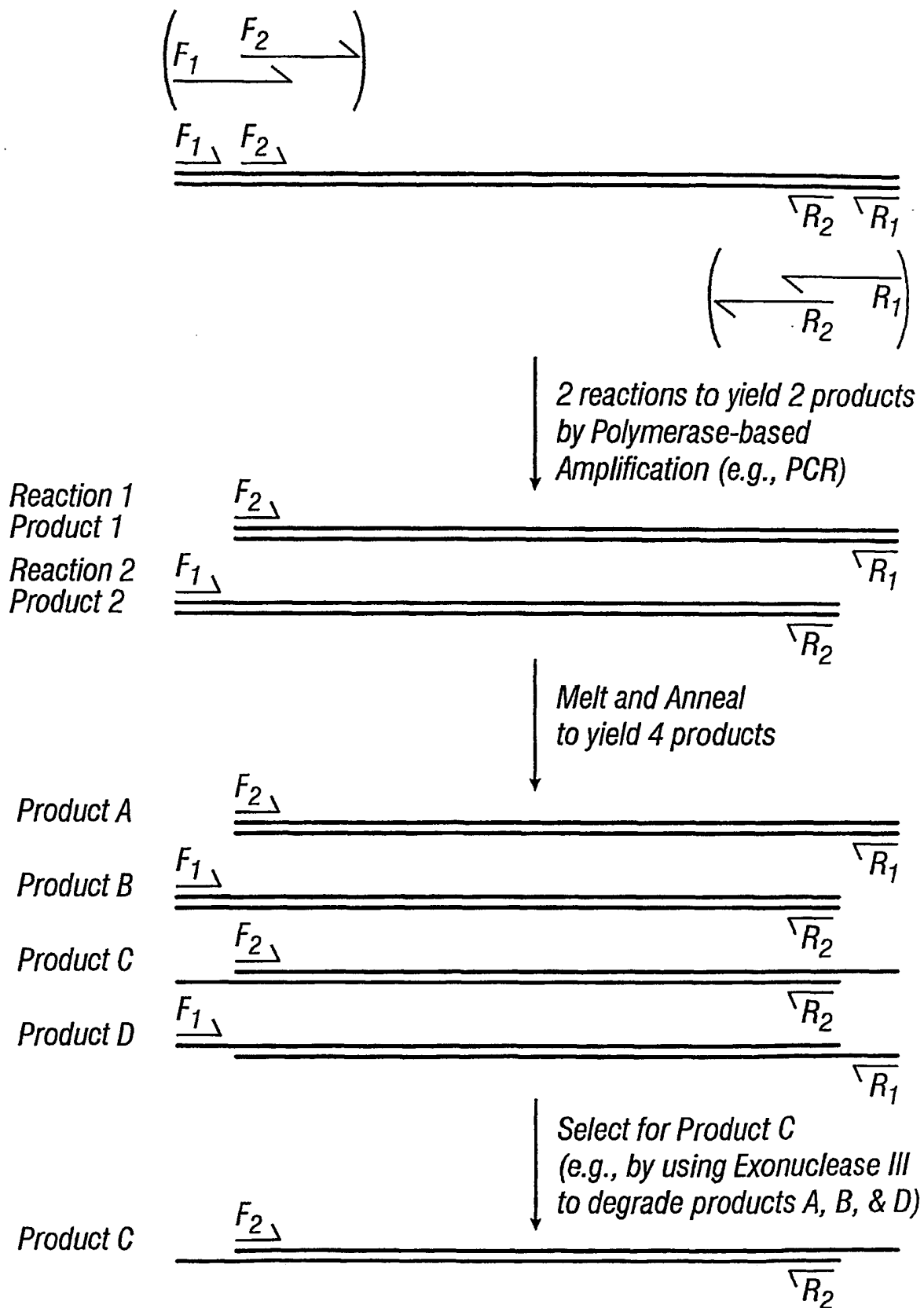
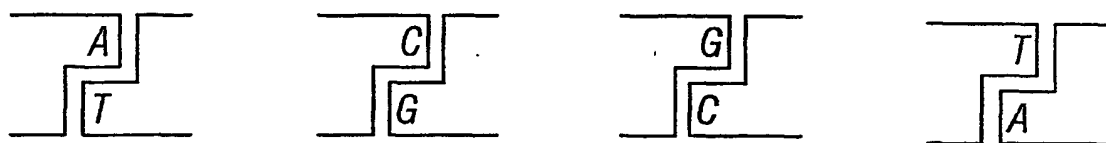


FIG. 2

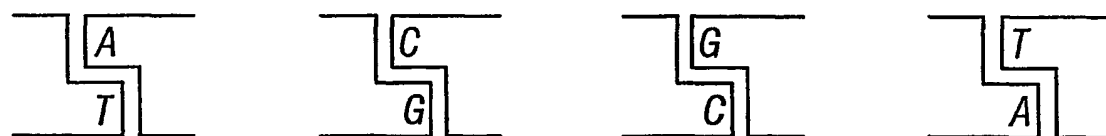
*Panel A.*



*Panel B.*



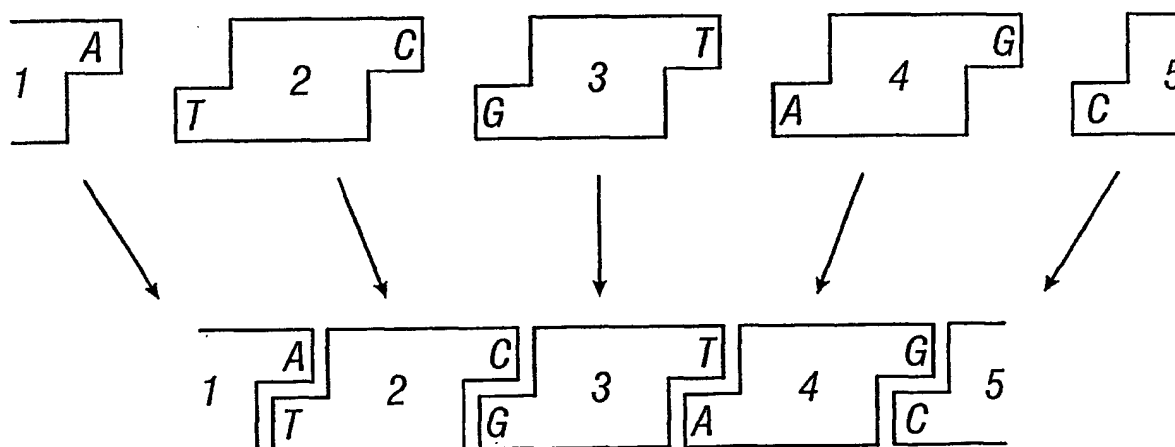
*Panel C.*



*Panel D.*



**FIG. 3**

*Panel A.**Panel B.***FIG. 4A**

Panel C.

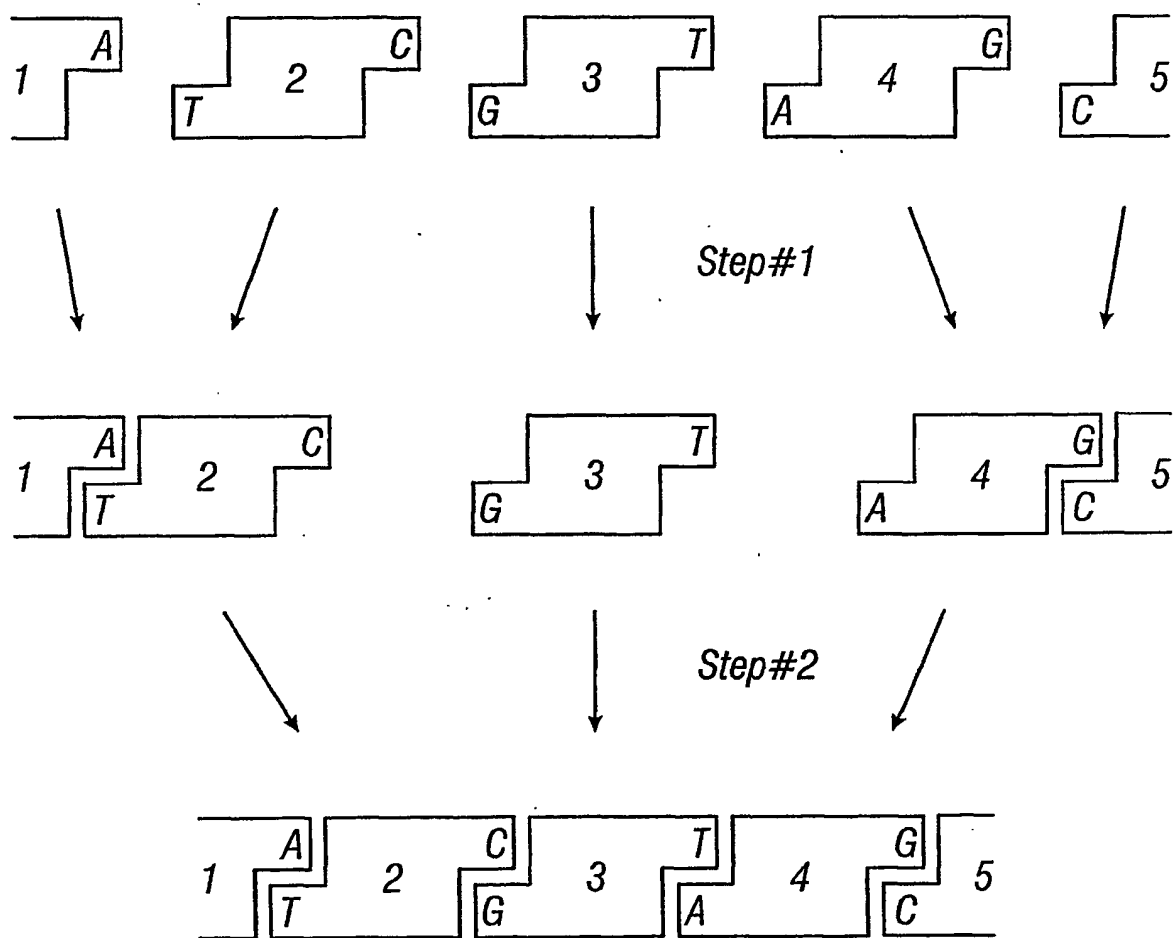


FIG. 4B



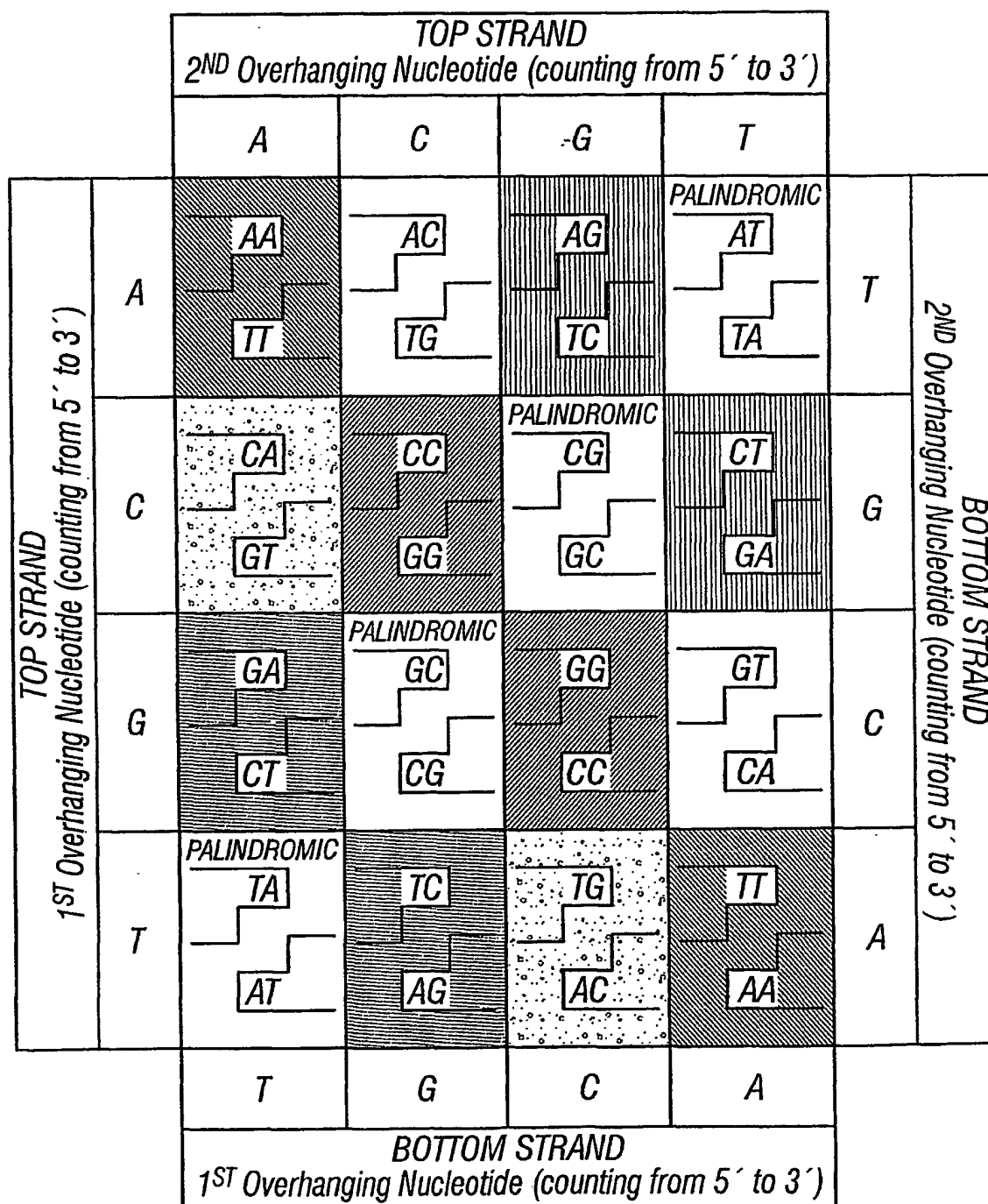
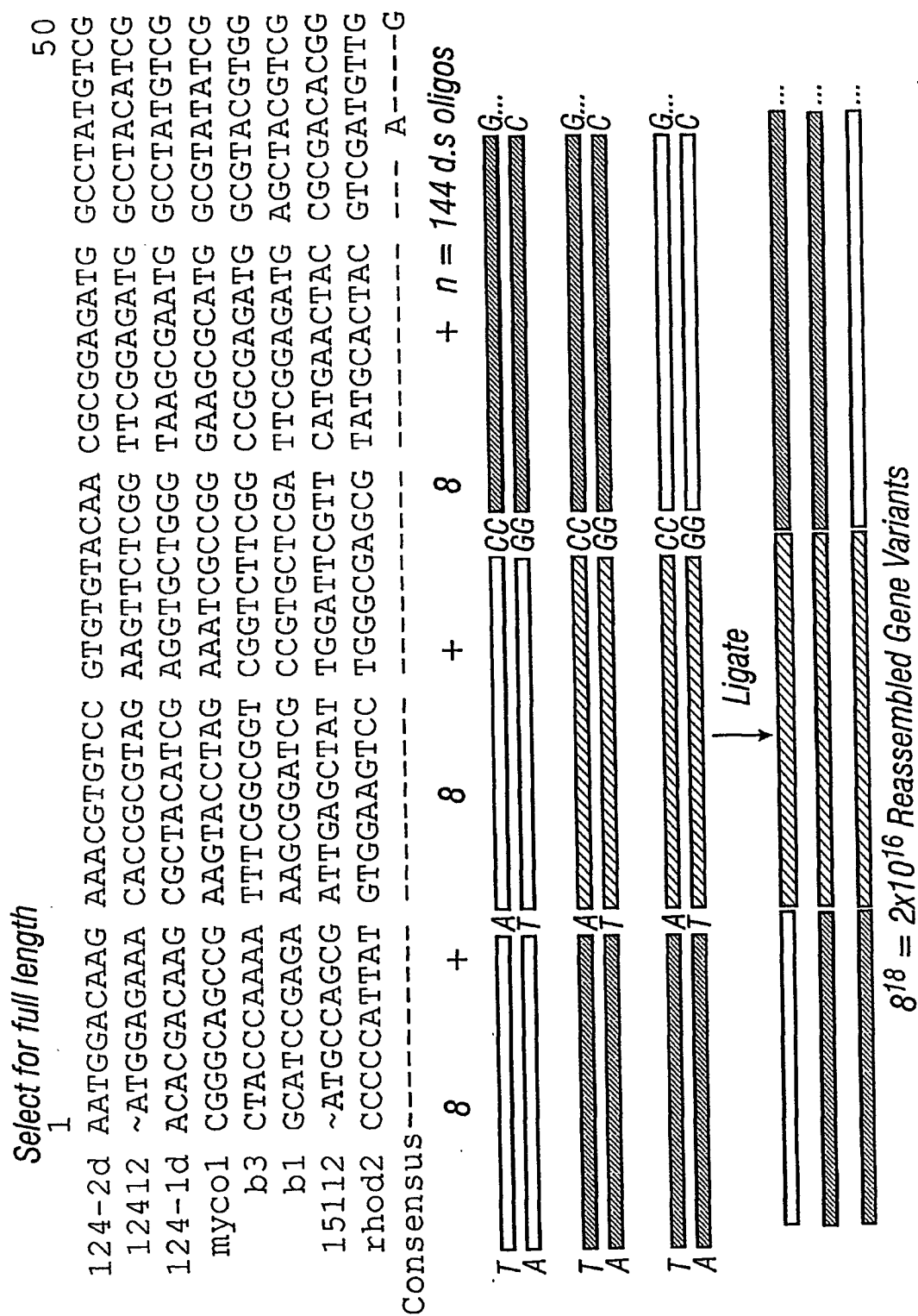


FIG. 5



100  
 ACACGGGCCA GGGTGATTCC GTTCTGTTTC TTCACGGCAA CCCGACGTCTG  
 ACGTGGGAGA GGGGACCCG ATCGTGTTCC TCCACGGAAA TCCACGTCTG  
 AGATGGGCGA GGTGATCCC ATCATTTTCC AACACGGCAA TCCGACCTCA  
 ACGAAGGCAA GGTGACGCC ATCGTCTTTC AGCACGGCAA CCCACGTCTG  
 AAGTGGGACG GGGGACCCC ATCGTACTCT TGCACGGCAA CCCACCTCTG  
 ATACCGGCGA GGGAGGCCG ATCGTGTTCC TTCACGGCAA CCCGACTTCC  
 GCGTCGGCGA T...CTTCCC GTCGTGTTCC TGCACGGCAA CCCACGTCTG  
 GACCGCGGGA TGGCACGCC TGTCTGTTCC TGCACGGTAA CCCGACCTCTG  
 -----G--- -T---T---T--- --CAcGG-AA -CC-AC-TC-

FIG. 6B

Represents 15% of gene

150  
 TCGTATCTGT GGAGGGGCGT AATGCCCTTT GTGACGGACG TCGCCCGATG  
 TCGTACCTGT GCGGGAACGT GATCCCCAC GTTGCCGGCT TGGACGCTG  
 TCGTACCTGT GCGCAACAT CATGCCCAT GTGCAACAGC TCGTCCGCTG  
 TCTTACTTGT GCGCAACAT CATGCCGCAC TTGGAAGGCG TGGCCCGGCT  
 TCGTACCTCT GCGCAACGT GTTGCCGCAC CTGGCGCCGT TAGGCCGCTG  
 TCTATCTTT GCGCAACAT CATCCCCAT CTCGCGGATC ACGCAGATG  
 TCTACGTCT GCGCAACGT GATCCCCGCAC GTCGCTGGCC AGCACCGGTG  
 TCTACCTGT GCGCAACAT CATCCCCCAT GTAGCACCGA GTCATCGGTG  
 TC--A--T-T GG-G---C-T --T-CC----- -T-----G-----

FIG. 6C

150am13_00	NCOI	<u>CA</u> TCATGCACG	CGGATATTTC	ATCGAGCAAT	GACACGGTCG	GCGTTG	CCGT
150AM7_001		<u>CA</u> TCATCACG	CGGACATTTC	ATCGAGCAAT	GACACGGTCG	GCGTTG	CCGT
431am7_002		<u>CA</u> TGAGACACG	GAGATATCTC	CAGCAGCAAC	GATTGCGTGG	GCGTGG	CCGT
150am13_00		CGTGAACTAC	AAGATGCCCTC	GCCTTCATAC	CAAGCGGAG	GT	TTAGCGA
150AM7_001		CGTGAACTAC	AAGATGCCGC	GGCTTCACAC	CAAGGCTGAG	GT	GCTGGCCA
431am7_002		CGTGAACTAC	AAGATGCCGC	GGCTGCATAC	CCGCGCGGAG	GT	GATGGAGA
150am13_00		ACGCCAGAAA	GATCGGCGAG	ATGATCGTCG	GCATGAAGAC	CGG	CCTGCCC
150AM7_001		ACTGCCGCAA	GATCGCCGAC	ATGCTGGTCG	GCATGAAGAG	CGG	CCTGCCC
431am7_002		ACGCCCGCAA	GATCGCCGAC	ATGCTCGTGG	GCATGAAGCG	CGG	CCTGCCC
150am13_00		GGAAATGGATC	TGGTGATCTT	CCCGGAATAT	TCGAC	CCACG	GCATCATGTA
150AM7_001		GGAAATGGATC	TGGTGATCTT	CCCGGAATAT	TCCAC	CCACG	GCATCATGTA
431am7_002		GGCATGGACC	TGGTCATCTT	CCCCGAGTAC	TCCAC	CCACG	GCATCATGTA
150am13_00		CGACTCCAAG	GAAATGTACG	ATACCGCGTC	CGTCGTGCC	GG	CGAGGAGA
150AM7_001		CGACTCCAAG	GAGATGTACG	ACACGGCGTC	GACGGTGCCG	GG	TGAAGAGA
431am7_002		CGACGCCAAG	GAAATGTACG	AAACCGCTTC	GGCCATTCCG	GG	CGAAGAGA
150am13_00		CCGAGATTTT	TGCCGAAGCC	TGCCGCAAGG	CGAAAAGTCTG	G	GGCGGTGTC
150AM7_001		CCGAGATTTT	CGCCGAGGCC	TGCCGCAAGG	CCAAGGTCTG	GG	CGGTGTC
431am7_002		CTGCTGTGTT	CGCCGACGCC	TGCCGCAAGG	CCAACGTATG	GG	CGGTGTT

FIG. 7A

150am13_00	TCGCTCACCG	GCGAACGTCA	CGAGGAACAT	CCGAAGAAAG	AAAG C	CGCCCTACAA
150AM7_001	TCGCTGACCG	GCGAGCGCCA	CGAGGAGCAT	CCCAATAAAG		CGCCGTACAA
431am7_002	TCGCTGACGG	GCGAGCGCCA	CGAAGAGCAC	CCGAACAAGG		CGCCGTACAA
					CAG AA	
150am13_00	CACGCTGATC	CTGATGAACG	ACAAGGGCGA	GGTGGTCCAG		AAATACCGCA
150AM7_001	CACCCTGATC	CTGATGAACG	ACAAGGGTGA	AGTCGTTT		AAATATCGCA
431am7_002	CACGCTCATC	CTGATGAACA	ACAAGGGCGA	GATCGTG		AAATACCGCA
				GGTA		
150am13_00	AGATCATGCC	GTGGGTTC	ATCGAGGGCT	GGTACCCCGG		CAACTGCACC
150AM7_001	AGATCATGCC	GTGGGTGCCG	ATCGAAGGCT	GGTATCCCGG		CAACTGCACG
431am7_002	AGATCATGCC	CTGGGTGCCG	ATCGAAGGCT	GGTATCCCGG		CGATTGCACC
			TGAAG			
150am13_00	TACGTCTCCG	ACGGGCCGAA	GGGCATGAAG	GTTTCGCTGA		TCATCTGCCA
150AM7_001	TACGTCTCCG	AAGCCCCGAA	GGGCATGAAG	ATGTCGCTGA		TCATCTGCCA
431am7_002	TATGTGTCCG	AAGCCCCCAA	GGGACTGAAG	ATCAGCCTCA		TCATCTGCCA
			TCTGGCG			
150am13_00	TGACGGCAAC	TATCCGGAAA	TCTGGCGCGA	CTGCGCCATG		AAGGGCGCCG
150AM7_001	CGACGGCAAC	TACCCGGAAA	TCTGGCGTGA	CTGCGCGATG		AAGGGCGCCG
431am7_002	CGACGGCAAT	TACCCCGAGA	TCTGGCGCGA	TTGCGCCCATG		CGCGGCGCCG
		CCAG				
150am13_00	AGCTGATCGT	GCGCTGCCAG	GGCTACATGT	ATCCGGCCAA		GGACCAGCAG
150AM7_001	AACTGATCAT	CCGCTGCCAG	GGCTACATGT	ATCCCGCCAA		GGATCAGCAG
431am7_002	AGCTGATCGT	GCGTTGCCAG	GGATACATGT	ACCCGGCCAA		GGACCAGCAG

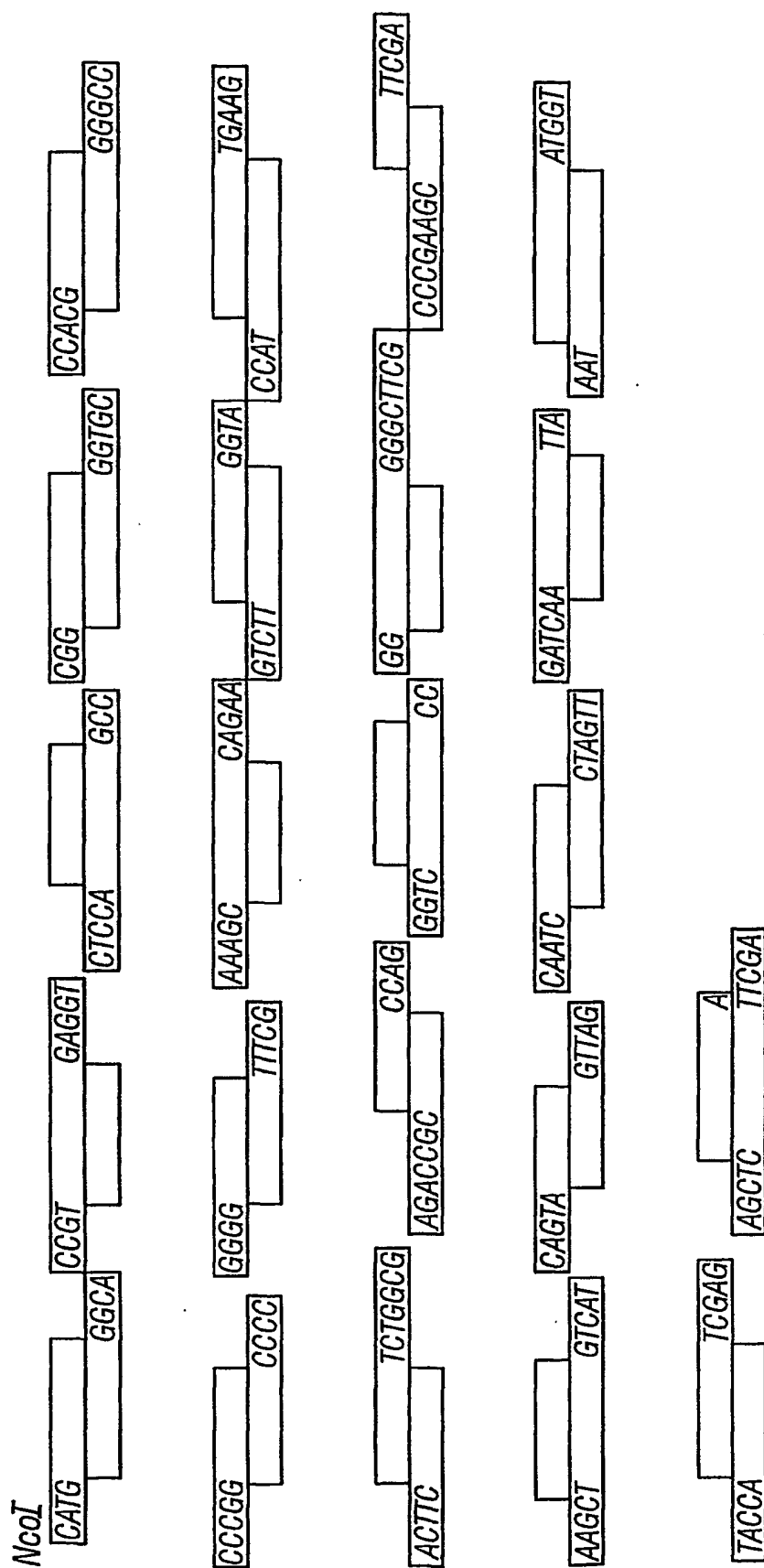
FIG. 7B

150am13_00	GC	GTCAATCATGG	CGAAGGC	GAT	GGCGTGGCGG	AATAATTGTT	ACGTCCGCGGT
150AM7_001		GTGCTGATGG	CGAAGCAAT		GGCCTGGGCC	AACAACGTTT	ATGTCGCGGT
431am7_002		GTCAATGGTGT	CCAAGGC	CAT	GGCGTGGATG	AACAACGTCT	ACGTGGCGGT
			GGGCTTCG				
150am13_00		TTCCAATGCC	GGGGCTTCG		ATGGCGTCTA	TTCGTATTTC	GGCCACTCGG
150AM7_001		CGCCAATGCC	TCGGGCTTCG		ACGGCGTCTA	CTCGTATTTC	GGCCATTTCG
431am7_002		GGCCAATGCC	GGGGCTTCG		ACGGCGTGTA	TTCCTACTTC	GGCCATTTCG
			TTCGA				
150am13_00		CGATCATCGG	CTTCGATGGC		CGCACGCTCG	GCGAATGCGG	CGAGGAAGAA
150AM7_001		CGATCATCGG	CTTCGACGGC		CGTACCCCTCG	GCGAATGCGG	CGAGGAGGAT
431am7_002		CCATCATCGG	CTTCGACGGC		CGCACGCTGG	GCGAATGCGG	TGAAGAAGAC
			C AGTA				
150am13_00		TACGGCATCC	AGTATGCCCA		GCTTTCGAAG	ATGCTGATCC	GCGACGCCCG
150AM7_001		TATGGCATCC	AGTATGCCCG		CATCTCCAAG	TCGCTGATCC	GCGACGCCCG
431am7_002		ATGGGCGTGC	AGTACGCCGA		GCTCTCCACC	AGCCTGATCC	GCGACGCCCG
			CAATC				
150am13_00		CCGCACCCGA	CAATCGGAAA		ACCATCTCTT	CAAGCTGGTG	CATCGTGGCT
150AM7_001		CCGCACCCGC	CAATCGGAAA		ACCATCTCTT	CAAGCTGGTG	CACCGTGGCT
431am7_002		CAAGAACATG	CAGTCGCAGA		ACCACTTGTT	CAAGCTGGTG	CACCGCGGCT
			GATCAA				
150am13_00		ACACCGGGTT	GATCAACTCC		GGCGAGGGCG	ACCGGGTCT	CGCGGCCCTGT
150AM7_001		ACACCGGCAT	GATCAATTCC		GGCGAGGGCG	ACCGGGTGT	CGCGGCTTGC
431am7_002		ACACCGGCAA	GATCAATTCC		GGCGAAGAGG	CCACCGGCGT	CGCGGCATGC

FIG. 7C

150am13_00	TTA	CC	TTA	TGAGT	TCTACAACAA	ATGGATCGCC	GATCCGGAAG	GCACCCGCCA
150AM7_001		CCGTA	TGATT		TCTATTGAA	ATGGATCGCC	GATCCCGAGG	GTACACGCCA
431am7_002		CCGTA	CAACT		TCTACGCCAA	CTGGATCAAC	GATCCGGAGG	GCACGCCCAA
		ATGGT						
150am13_00		A	ATGGT	CGAG	TCCTTTACCC	GGCCGACGGT	GGGAACCGAT	GAAAGCGCCCA
150AM7_001		G	ATGGT	GGAA	TCCTTCACGC	GTCCGACGGT	GGGTGTGGAG	GAATGCCCCA
431am7_002		G	ATGGT	CGAA	TCCTTCACCC	GTCCACCCGT	GGGCACGCCG	GAGTGCCCCA
		TCGAG						
150am13_00		T	CGAA	GGCAT	CCCGAACAAAG	GTCCGGGTGC	ACCGCTGA	aagct
150AM7_001		T	CGAG	GGCAT	TCCGAACAAAG	GCCACCAACGC	ACCGCTGA	aagct
431am7_002		T	GGAC	GGCAT	CCCCAACGAG	GACGCCAAGC	ACCGCTAG	aagct
		HindIII						

FIG. 7D



*HindIII*

FIG. 8





FIG. 9

Gap Ligation

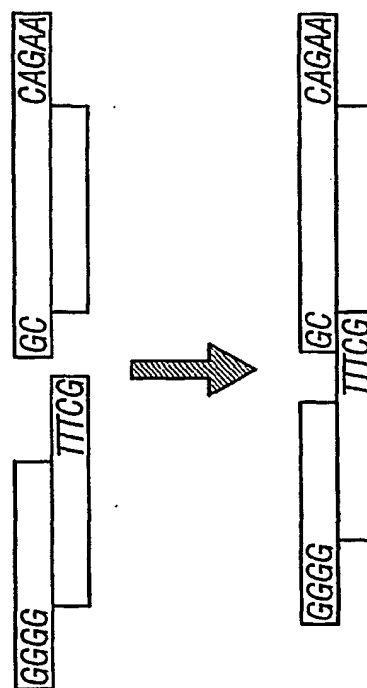
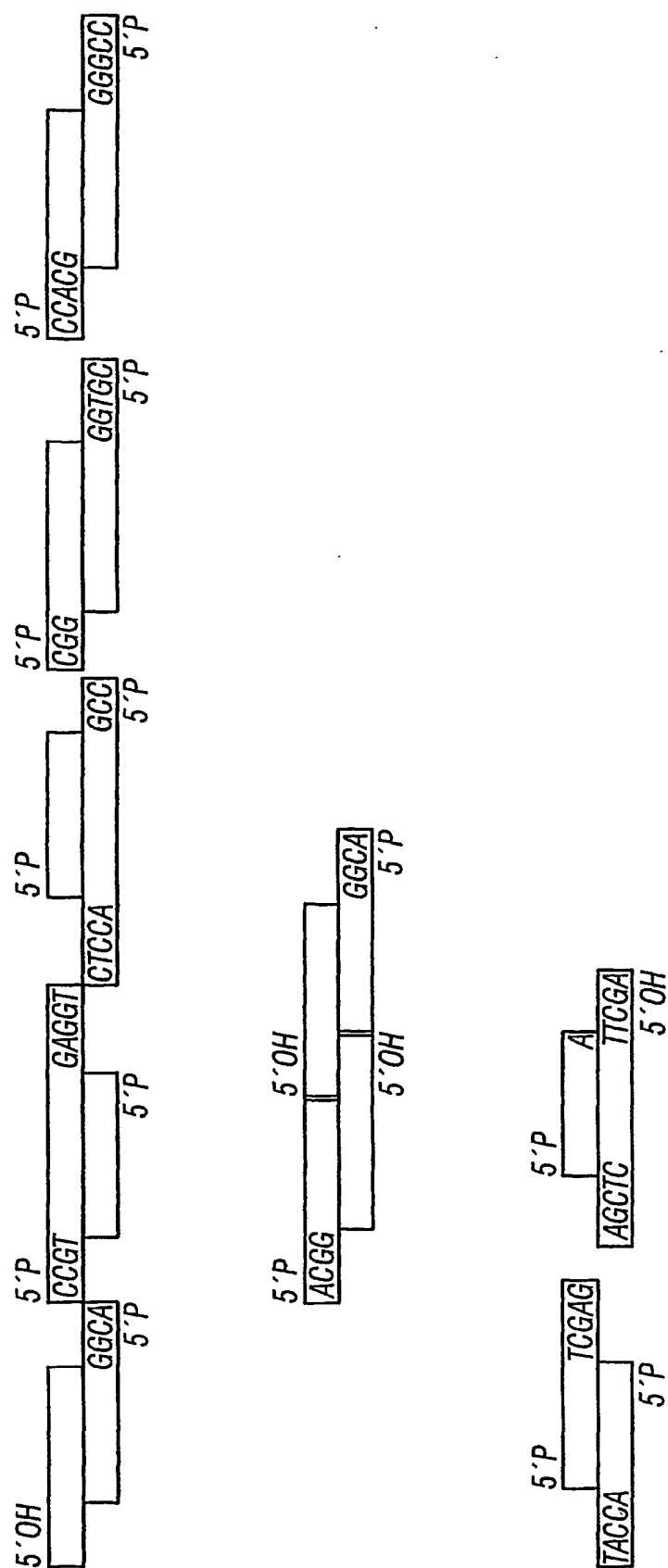


FIG. 10



**FIG. 11**

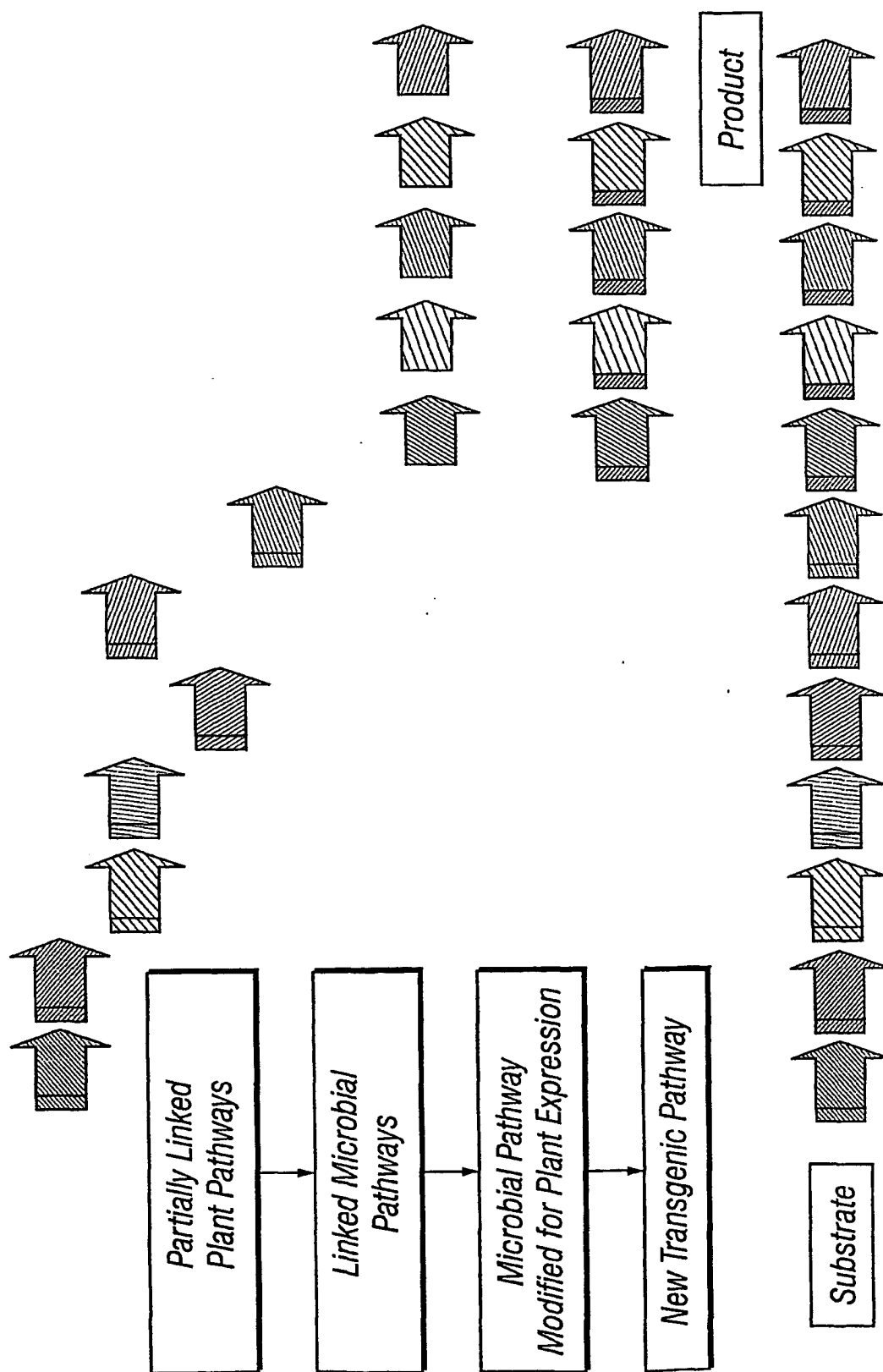


FIG. 12

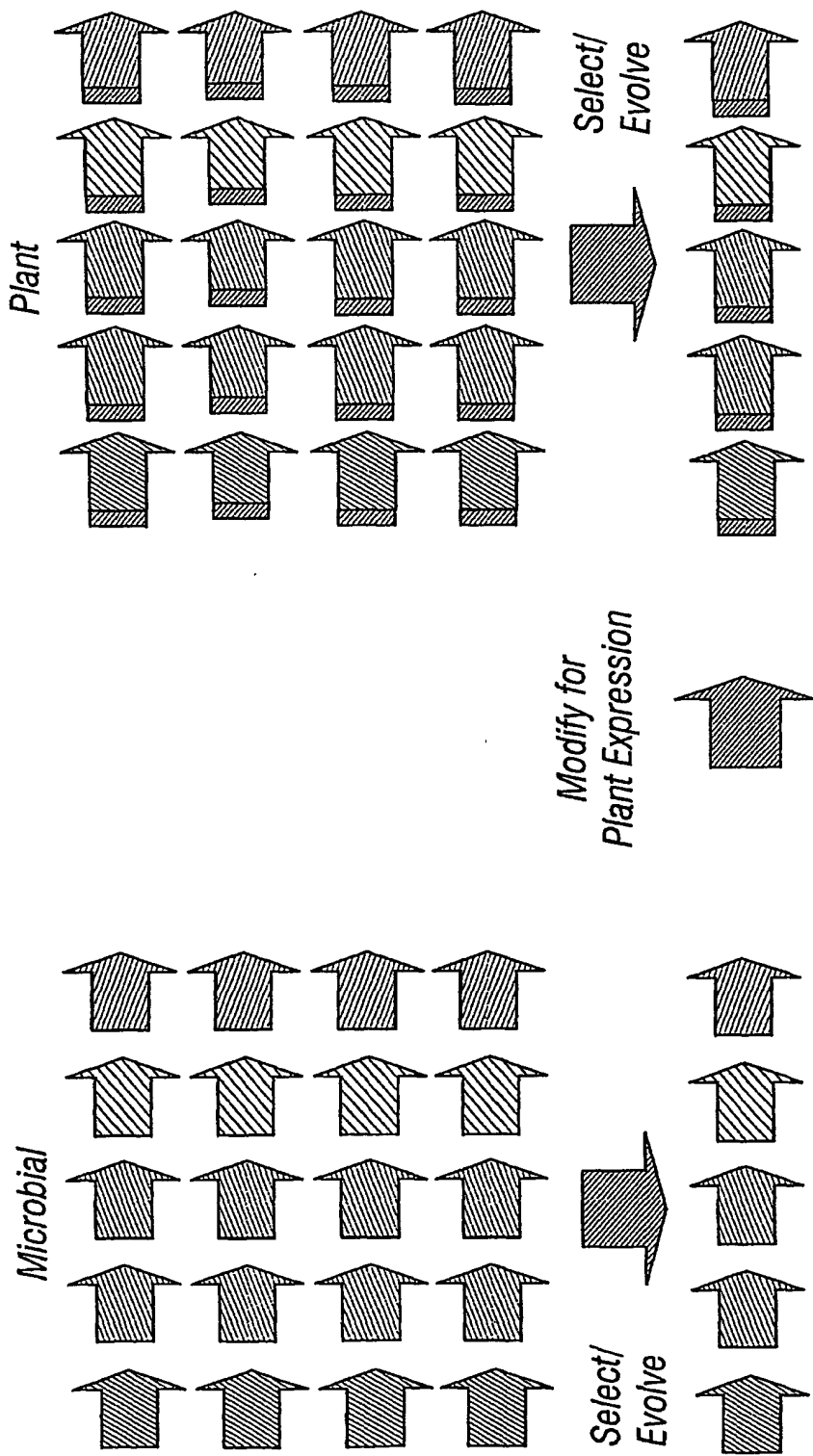


FIG. 13

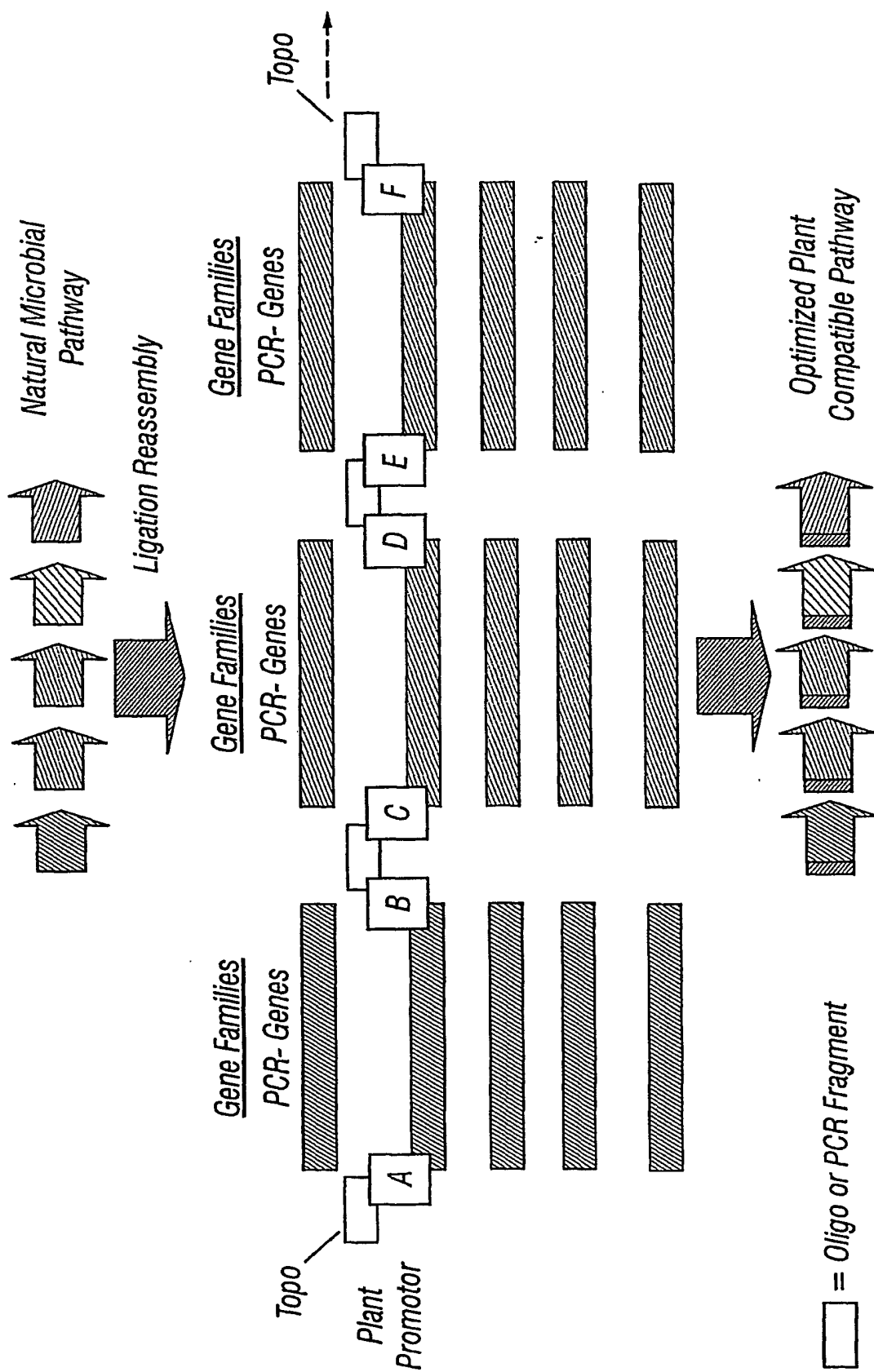
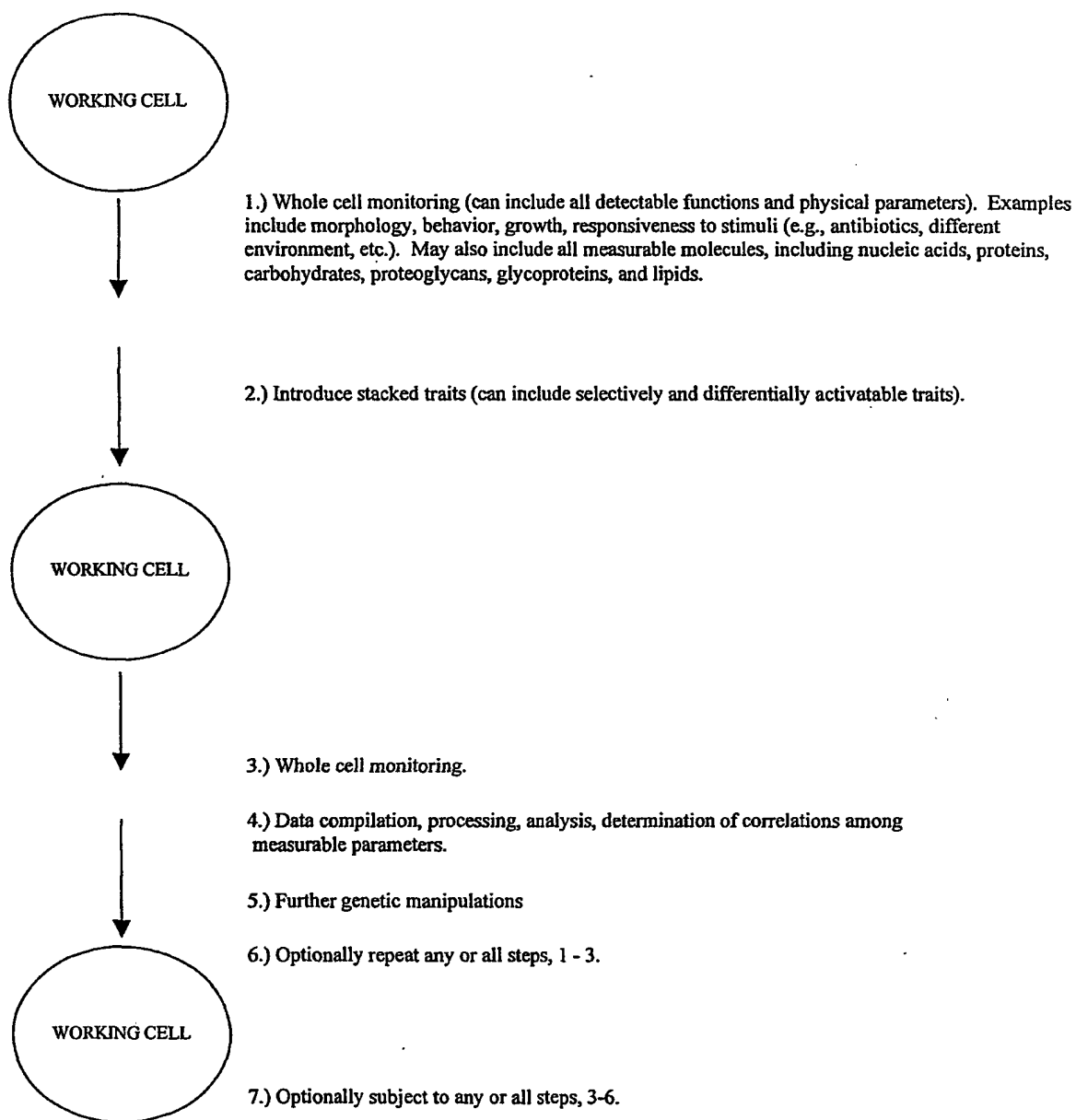


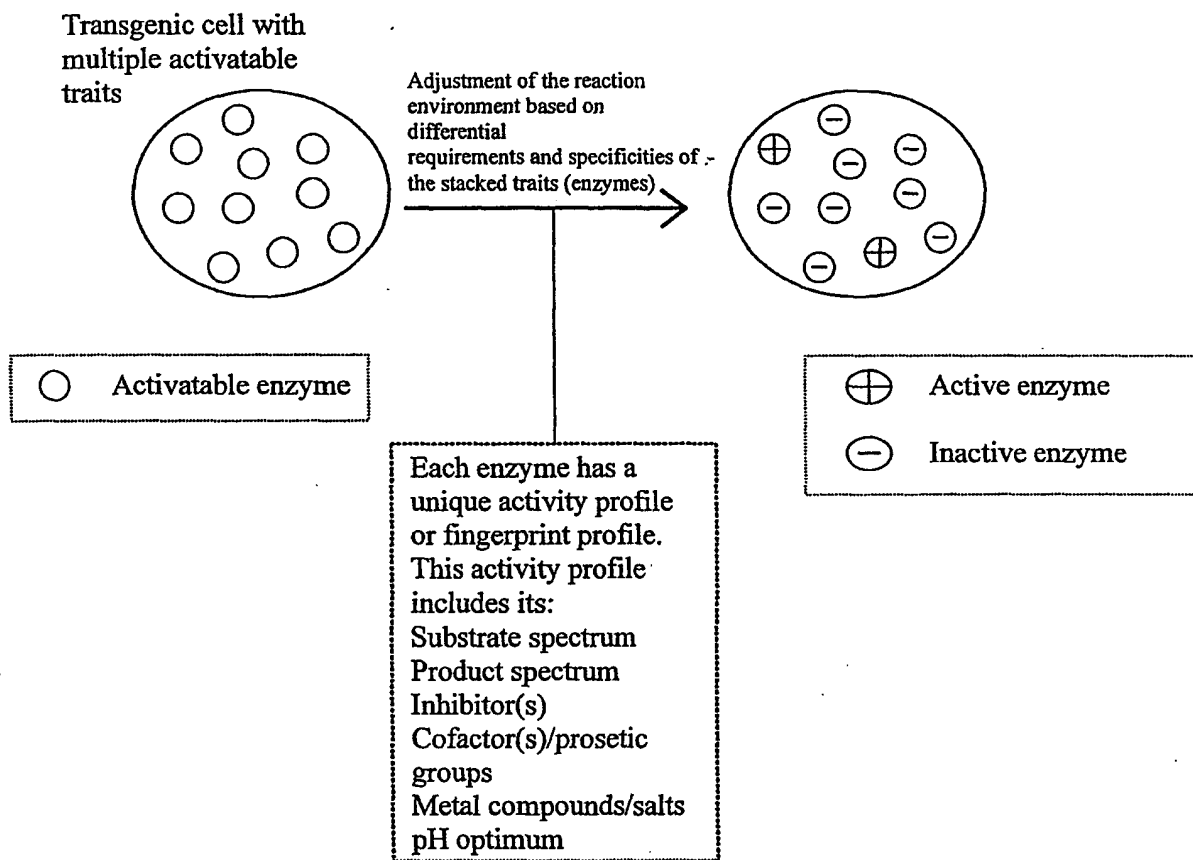
FIG. 14

**Fig. 15. HOLISTIC ENGINEERING OF DIFFERENTIALLY  
ACTIVATABLE STACKED TRAITS IN NOVEL TRANSGENIC  
PLANTS USING DIRECTED EVOLUTION AND WHOLE CELL  
MONITORING**



**Fig. 16. Differential Activation of Selected Traits Can Be Achieved by Adjusting and Controlling the Environment of the Traits.**

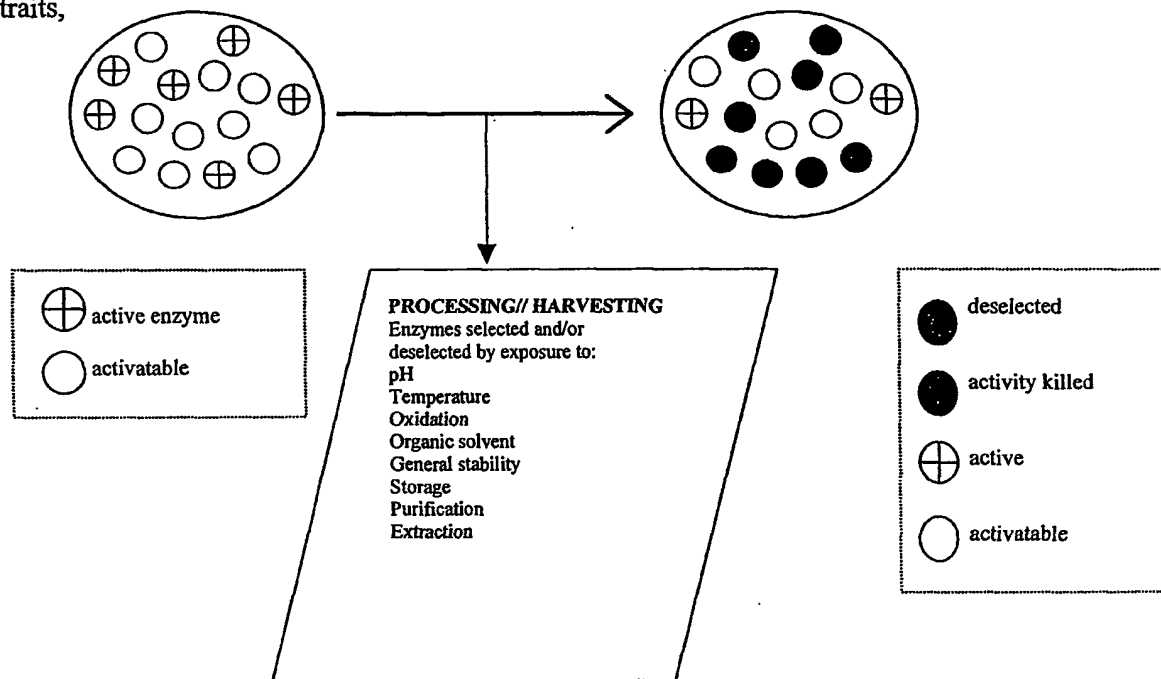
For example, in one aspect, stacked traits can be comprised of genetically introduced enzymes. Because the stacked enzymes have different activity profile (including reaction specificities and reaction requirements) they can be selectively and differentially activated by adjusting the environment to which they are exposed.



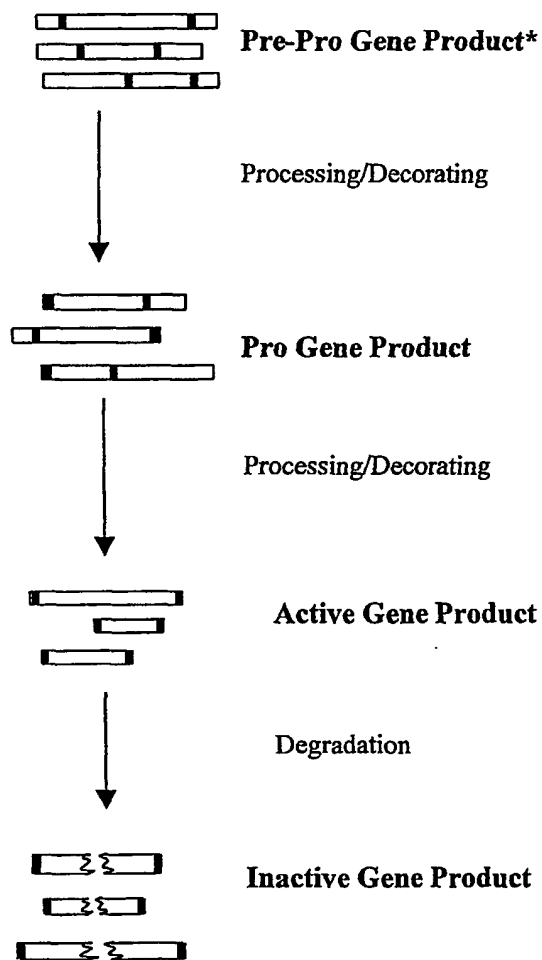
**Fig. 17. Harvesting, Processing, Storage**

Differentially activated and/or selected enzymes respond to the environments of harvesting, processing and storage to activate environmentally action specific promoters.

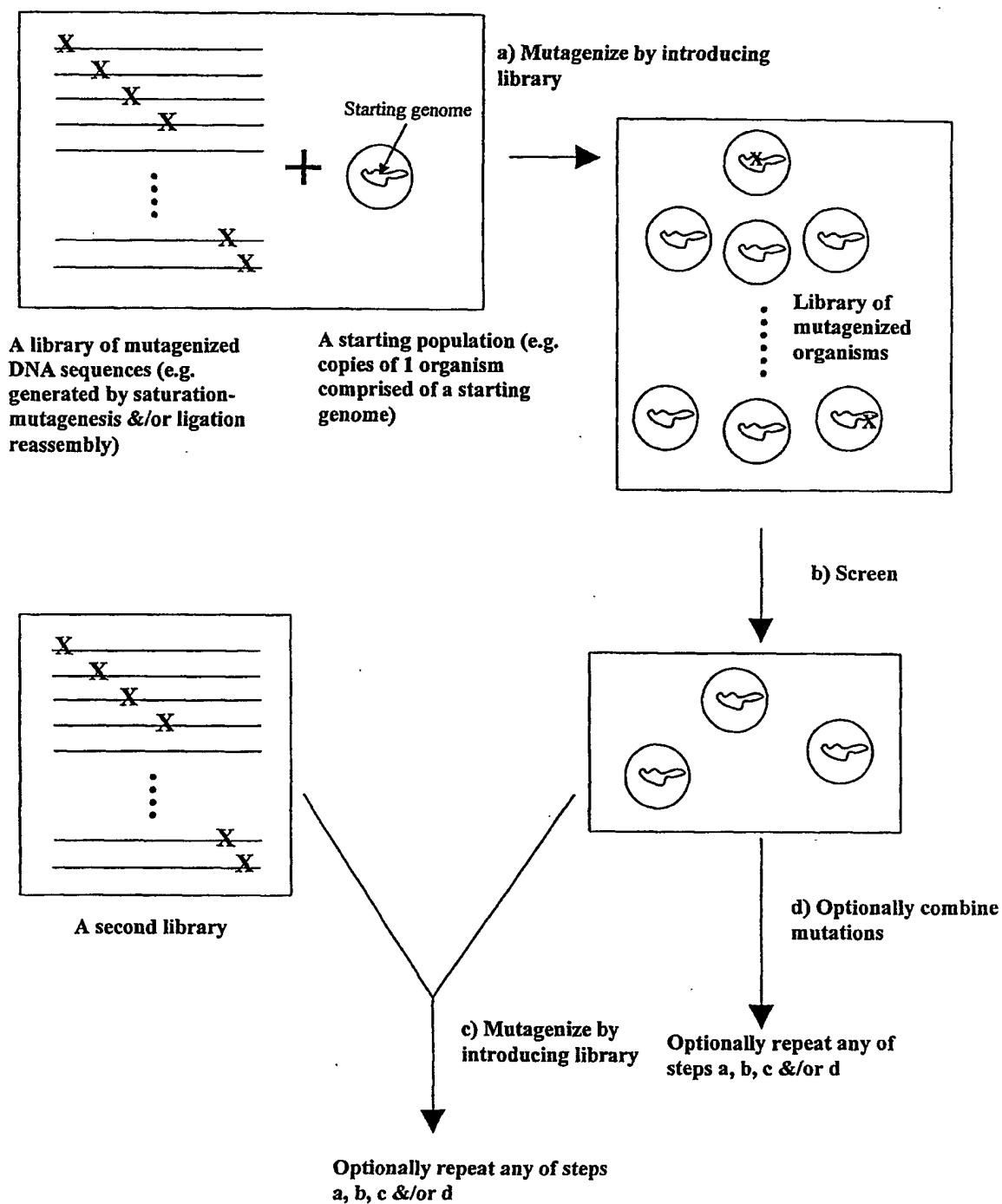
Transgenic cell with  
multiple activatable  
traits,





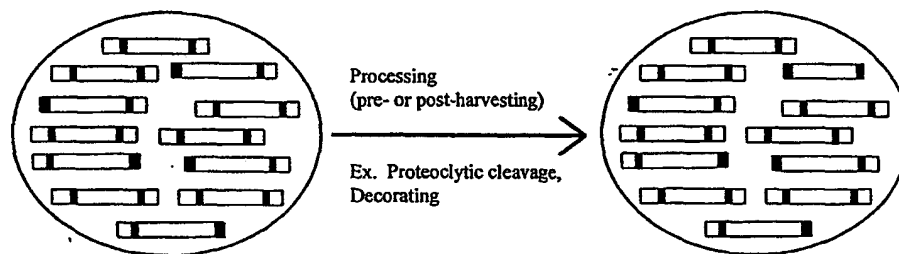
**Fig. 18. Processing**

\* An example of a Gene Product might be a protein. Through processing/decorating the protein changes forms, eventually becoming active. It is at this point that specific traits can be expressed differentially.

**Fig. 19. Cellular Mutagenesis.**

**Fig. 20. Differential Activation of Selected Precursor (Inactive) Gene Products**

Differential activation of selected precursor (inactive) gene products by controlling the post-translational modifications that differentially transform selected molecules from inactive precursor form to active form. Deselection of particular molecules can also be achieved by degradation (ex. By proteolytic cleavage).



Inactive precursor gene products  
(ex. pre-pro hormones, pro-hormones  
pre-pro proteins, or pro-proteins).

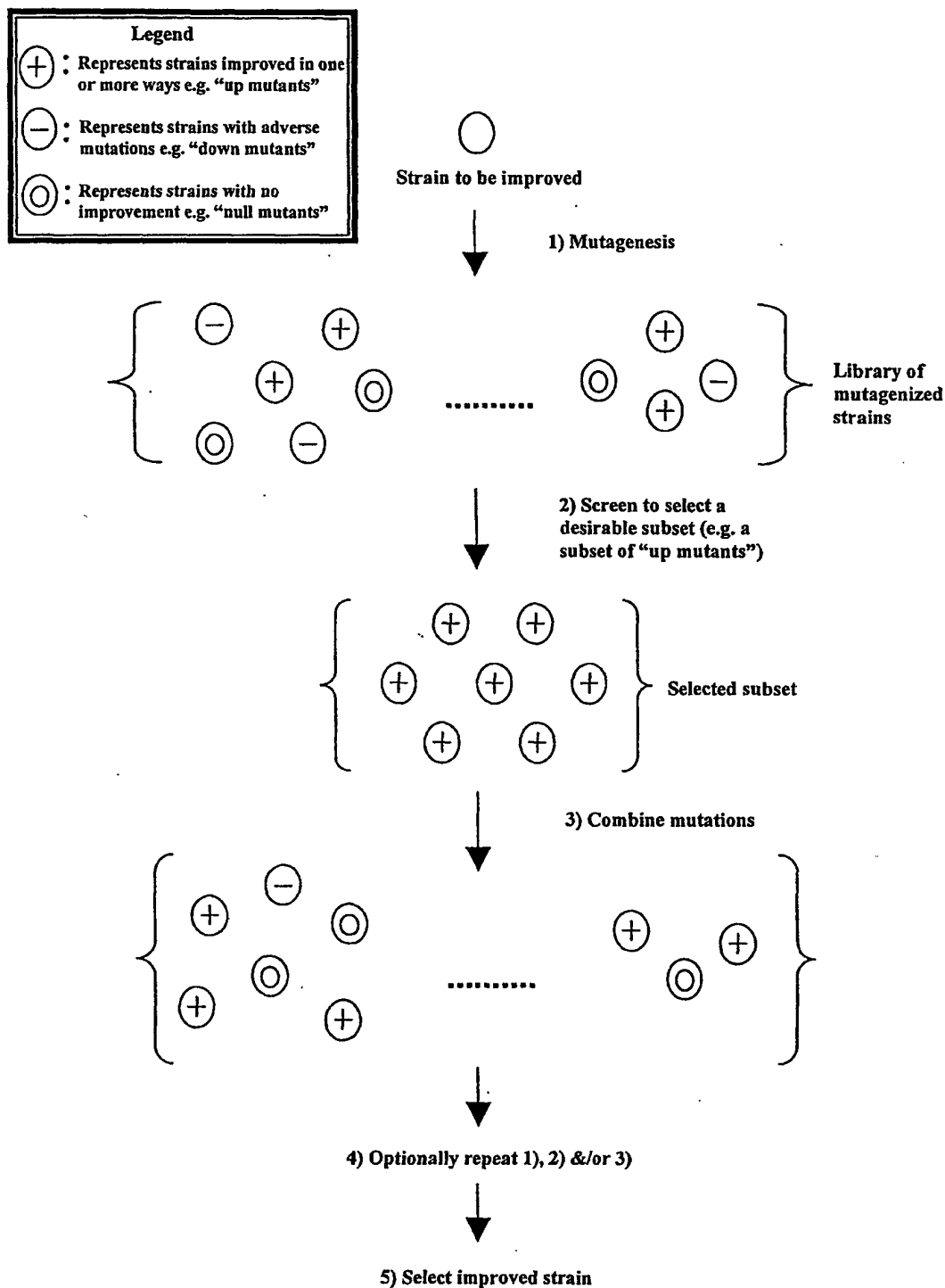
LEGEND:

 pre-pro

 pro

 active

**Figure 21. Starting population comprised of an organism strain to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved organism strain that has a desired trait**



**Figure 22. Starting population comprised of a genomic sequence to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved genomic sequence that has a desired trait**

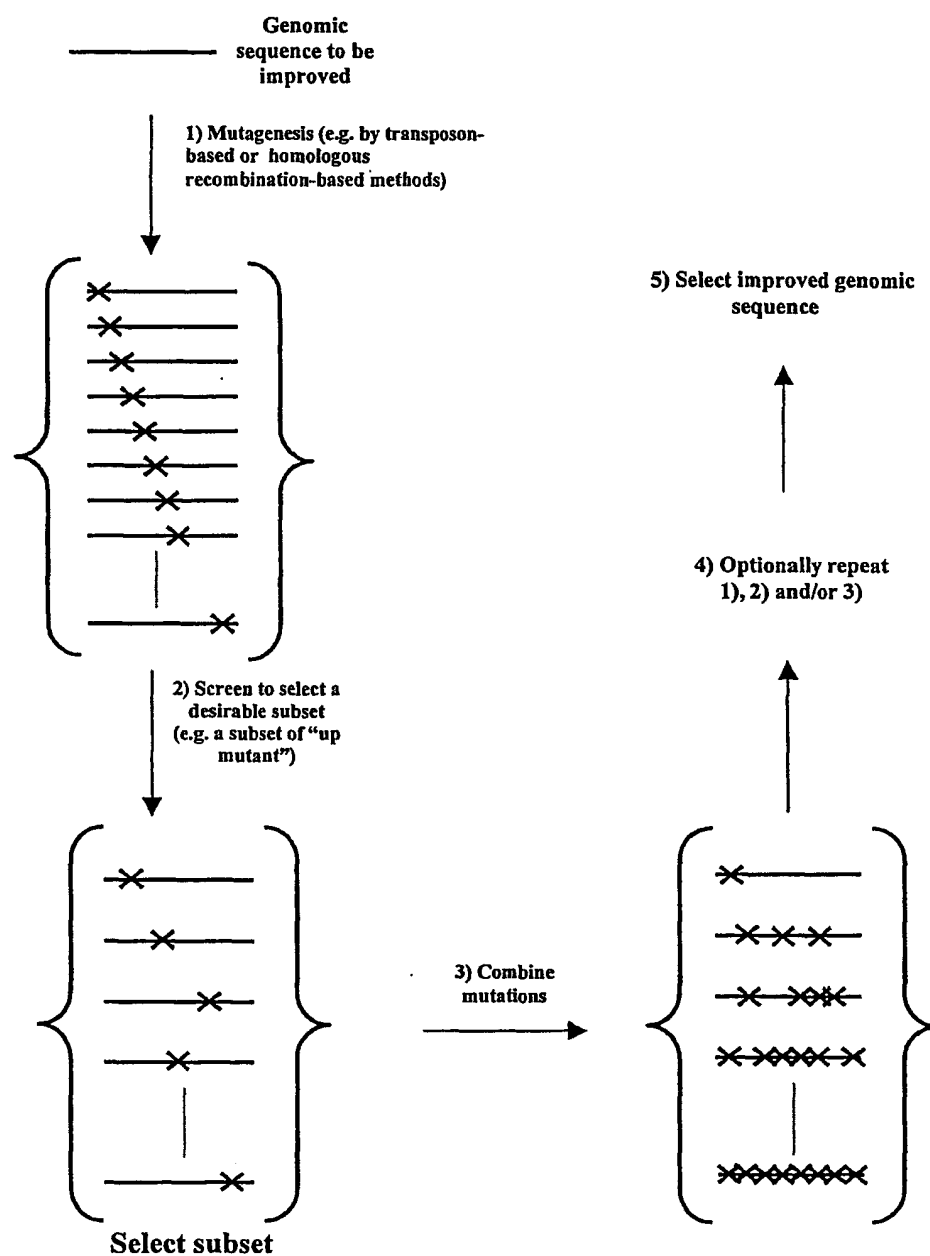


Fig. 23. Strain Improvement.

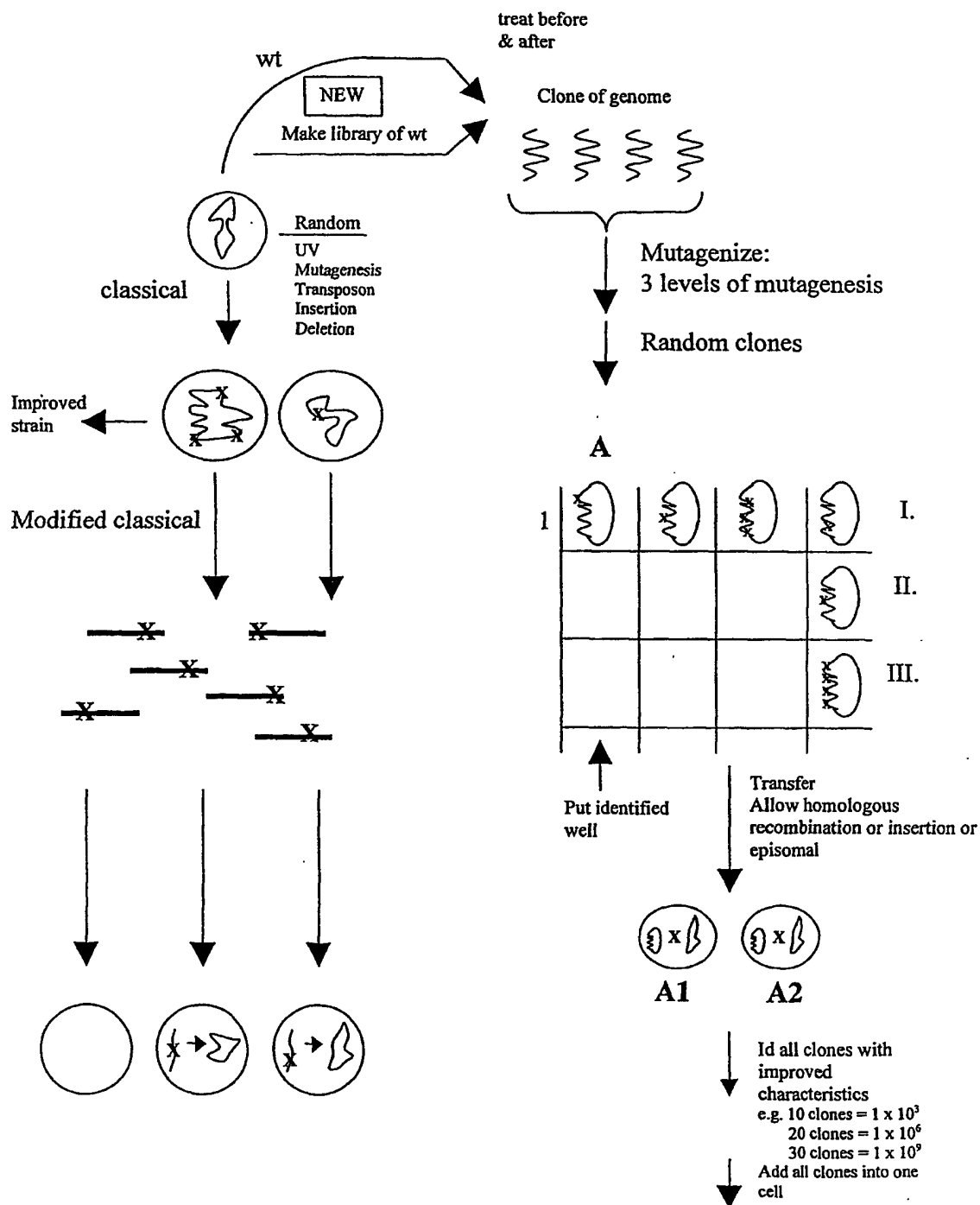
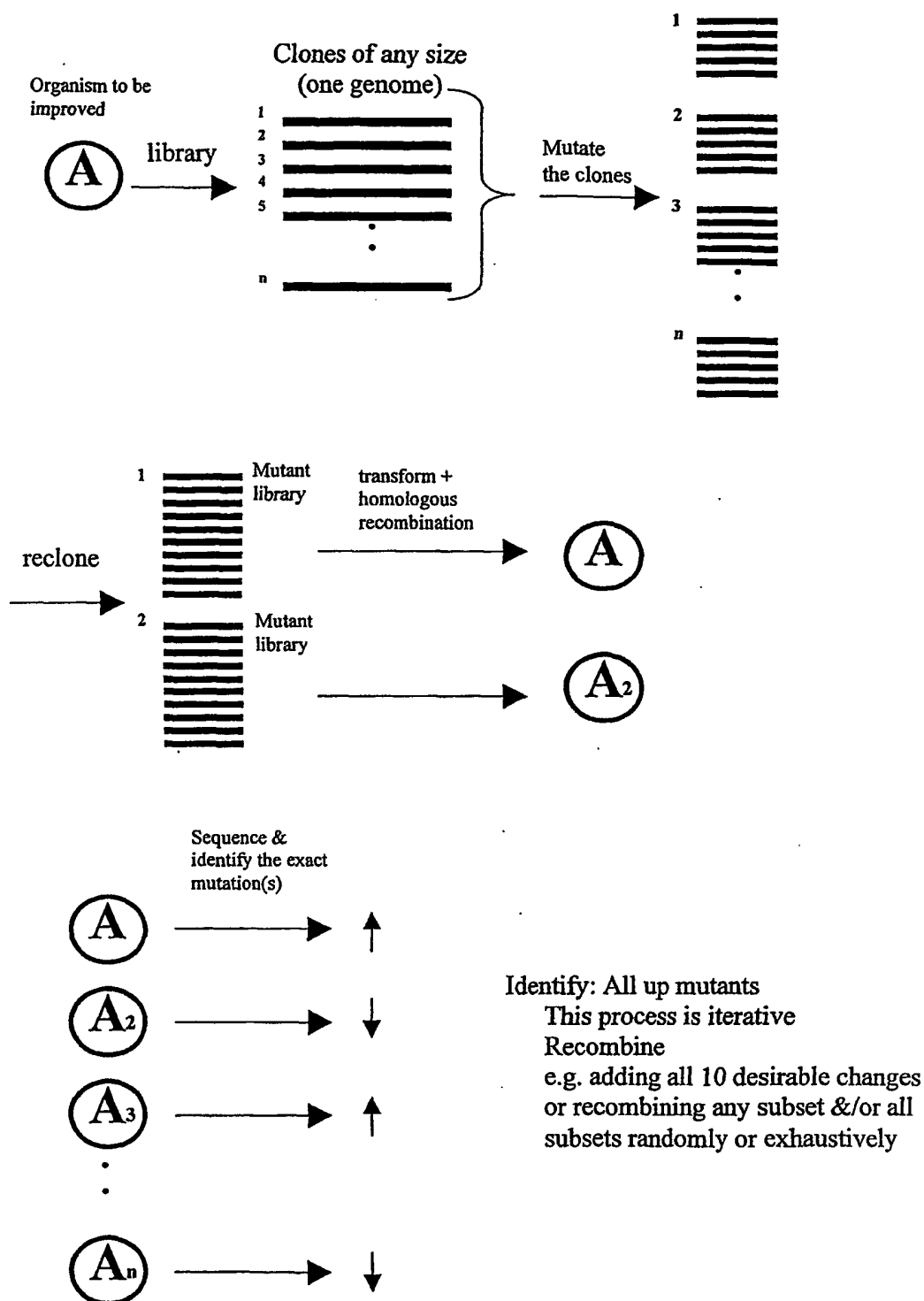


Fig. 24. Iterative Strain Improvement.



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
20 December 2001 (20.12.2001)

PCT

(10) International Publication Number  
**WO 01/96551 A3**

- (51) International Patent Classification<sup>7</sup>: C12N 15/10, C12Q 1/68
- (21) International Application Number: PCT/US01/19367
- (22) International Filing Date: 14 June 2001 (14.06.2001)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:  
09/594,459 14 June 2000 (14.06.2000) US  
09/677,584 30 September 2000 (30.09.2000) US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:  
US 09/594,459 (CIP)  
Filed on 14 June 2000 (14.06.2000)  
US 09/677,584 (CIP)  
Filed on 30 September 2000 (30.09.2000)
- (71) Applicant (for all designated States except US): **DI-VERSA CORPORATION** [US/US]; 4955 Directors Place, San Diego, CA 92121 (US).
- (72) Inventor; and
- (75) Inventor/Applicant (for US only): **SHORT, Jay, M.** [US/US]; 6801 Paseo Delicias, P.O. Box 7214, Rancho Santa Fe, CA 92067-7214 (US).
- (74) Agent: **HAILE, Lisa, A.**; Gray Cary Ware & Freidenrich LLP, Suite 1100, 4365 Executive Drive, San Diego, CA 92121-2133 (US).
- (81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EF, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:**  
— with international search report
- (88) Date of publication of the international search report:  
23 May 2002
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

WO 01/96551 A3

(54) Title: WHOLE CELL ENGINEERING BY MUTAGENIZING A SUBSTANTIAL PORTION OF A STARTING GENOME, COMBINING MUTATIONS, AND OPTIONALLY REPEATING

(57) Abstract: An invention comprising cellular transformation, directed evolution, and screening methods for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable. Also, a method of retooling genes and gene pathways by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, thus conferring operability to a novel gene pathway when it is introduced into an intended host. For example a novel man-made gene pathway, generated based on microbially-derived progenitor templates, that is operable in a plant cell. Furthermore, a method of generating novel host organisms having increased expression of desirable traits, recombinant genes, and gene products.



## INTERNATIONAL SEARCH REPORT

International Application No

PC1/US 01/19367

A. CLASSIFICATION OF SUBJECT MATTER  
IPC 7 C12N15/10 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 C12N C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

WPI Data, PAJ, CAB Data, SEQUENCE SEARCH, BIOSIS, EPO-Internal

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>HUTCHISON C A ET AL: "GLOBAL TRANSPOSON MUTAGENESIS AND A MINIMAL MYCOPLASMA GENOME"</p> <p>SCIENCE, AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE., US, vol. 286, no. 5447, 10 December 1999 (1999-12-10), pages 2165-2169, XP000865808</p> <p>ISSN: 0036-8075</p> <p>the whole document</p> <p style="text-align: center;">--- -/--</p>	<p>1-4, 8-10, 21</p>



Further documents are listed in the continuation of box C



Patent family members are listed in annex.

## \* Special categories of cited documents:

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

\*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

\*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

\*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

\*G\* document member of the same patent family

Date of the actual completion of the international search

10 December 2001

Date of mailing of the international search report

14/12/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Hornig, H

## INTERNATIONAL SEARCH REPORT

Int'l Application No

PCI/US 01/19367

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	"DEFINED OLIGONUCLEOTIDE TAG POOLS AND PCR SCREENING IN SIGNATURE-TAGGED MUTAGENESIS OF ESSENTIAL GENES FROM BACTERIA" BIOTECHNIQUES, EATON PUBLISHING, NATICK, US, vol. 26, no. 3, March 1999 (1999-03), pages 473-474, 476, 478-480, XP000938981 ISSN: 0736-6205 the whole document ---	1-4, 8-10, 21
X	HENSEL M ET AL: "SIMULTANEOUS IDENTIFICATION OF BACTERIAL VIRULENCE GENES BY NEGATIVE SELECTION" SCIENCE, AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE, US, vol. 269, 21 July 1995 (1995-07-21), pages 400-403, XP000645478 ISSN: 0036-8075 the whole document ---	1-4, 8-10, 21
P, X	WO 00 55346 A (PARADIGM GENETICS INC ;HAMER LISBETH (US); HAMER JOHN E (US)) 21 September 2000 (2000-09-21) the whole document ---	1-4, 8-10, 21
P, X	WO 01 02555 A (CAMACHO LUIS ;GICQUEL BRIGITTE (FR); GUILHOT CHRISTOPHE (FR); PAST) 11 January 2001 (2001-01-11) claims 1-35 ---	1-4, 8-10, 21
A	WO 00 18906 A (MAXYGEN INC ;LIU LU (US); PATTEN PHILLIP A (US); STEMMER WILLEM P) 6 April 2000 (2000-04-06) page 51, line 11; claims 1-51; figures 1-15 ---	22
A	WO 97 26334 A (HALBAN PHILIPPE A ;NEWGARD CHRISTOPHER B (US); NORMINGTON KARL D ( )) 24 July 1997 (1997-07-24) claims 1-112 ---	22
A	FR 2 761 576 A (INST NAT SANTE RECH MED) 9 October 1998 (1998-10-09) claims 1-19 ---	22
A	WO 00 04190 A (SUBRAMANIAN VENKITESWATAN ;HUISMAN GJALT (US); MAXYGEN INC (US); T) 27 January 2000 (2000-01-27) the whole document ---	
A	WO 98 31837 A (DELCARDAYRE STEPHEN B ;MAXYGEN INC (US); MINSHULL JEREMY (US); NES) 23 July 1998 (1998-07-23) the whole document -----	

## INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PC1/US 01/19367

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
WO 0055346	A	21-09-2000	AU 3900700 A WO 0055346 A2	04-10-2000 21-09-2000
WO 0102555	A	11-01-2001	AU 5559300 A WO 0102555 A1	22-01-2001 11-01-2001
WO 0018906	A	06-04-2000	AU 1199000 A EP 1117777 A2 WO 0018906 A2 AU 2415200 A AU 3210100 A AU 3210200 A EP 1151409 A1 EP 1062614 A1 EP 1072010 A2 EP 1130093 A1 EP 1108783 A2 EP 1108781 A2 WO 0042559 A1 WO 0042560 A2 WO 0042561 A2 US 6319714 B1	17-04-2000 25-07-2001 06-04-2000 01-08-2000 01-08-2000 01-08-2000 07-11-2001 27-12-2000 31-01-2001 05-09-2001 20-06-2001 20-06-2001 20-07-2000 20-07-2000 20-07-2000 20-11-2001
WO 9726334	A	24-07-1997	US 6087129 A AU 718254 B2 AU 1750597 A AU 1830997 A CA 2246268 A1 CA 2246431 A1 EP 0876484 A1 EP 0910578 A2 WO 9726334 A1 WO 9726321 A2 US 6110707 A US 6194176 B1	11-07-2000 13-04-2000 11-08-1997 11-08-1997 24-07-1997 24-07-1997 11-11-1998 28-04-1999 24-07-1997 24-07-1997 29-08-2000 27-02-2001
FR 2761576	A	09-10-1998	FR 2761576 A1 AU 7054698 A EP 0972018 A1 WO 9845422 A1	09-10-1998 30-10-1998 19-01-2000 15-10-1998
WO 0004190	A	27-01-2000	AU 5102699 A EP 1108058 A1 WO 0004190 A1 US 6287862 B1	07-02-2000 20-06-2001 27-01-2000 11-09-2001
WO 9831837	A	23-07-1998	AU 5920998 A EP 1007732 A1 JP 2001508662 T WO 9831837 A1 US 6251674 B1 US 6287862 B1	07-08-1998 14-06-2000 03-07-2001 23-07-1998 26-06-2001 11-09-2001